

# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

*A single body-size gene tuned  
to both male and female needs  
in Atlantic salmon* **PAGE 405**

## Gender gap

SCIENCE AND SOCIETY

### THE MYTH BUSTERS

*A mugget of truth is all it  
takes to sustain false beliefs*

**PAGE 322**

CHEMISTRY

### WHITHER TOTAL SYNTHESIS?

*The big, the beautiful  
and the useful*

**PAGE 327**

SCIENCE COMMUNICATION

### CHANGE THE WORLD

*A how-to guide for  
scientists in the public eye*

**PAGE 332**

**NATURE.COM/NATURE**

17 December 2015 £10

Vol. 528, No. 7582

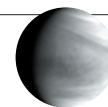


# THIS WEEK

## EDITORIALS

**REFUGEES** Germany shows the benefits to science and society **p.308**

**MENTAL HEALTH** The online mistakes that reveal global stigma **p.309**



**CLOSE-UP** Probe prepares to examine atmosphere on Venus **p.312**

## A seismic shift

*After 25 years of divisive debate, the governments of the world unite in Paris to fight global warming. But the hard work must start now.*

On 12 December, French foreign minister Laurent Fabius passed into existence a landmark agreement on global warming, and without a single word of discussion. The small green gavel produced only a soft crack at the United Nations climate summit in Paris, a sound quickly overwhelmed by a standing ovation. But that sound should echo. It ushered in a seismic shift in international environmental and economic policy. If everything goes according to plan, the reverberations will be felt around the world for decades — and perhaps centuries — to come.

The Paris agreement strengthens the previous goal of limiting warming to 2 °C above pre-industrial levels, ultimately suggesting that governments should “pursue efforts to limit the temperature increase to 1.5 °C”. Pushed by a coalition of island nations and some of the most vulnerable countries on Earth, this change offers a nod to scientific research, which suggests that even the 1 °C of warming experienced thus far is already having effects. Current commitments to reduce emissions might put the world on a path to keep the rise in temperature below 3 °C, and even that assumes substantial action in the decades to come. But all countries must revisit — and hopefully strengthen — their pledges every five years, beginning in 2020.

Despite the contradiction between commitments and goals, the Paris accord is a vast improvement over the last binding agreement to curb emissions. The 1997 Kyoto Protocol explicitly divided the world into two factions, rich and poor, and it required only rich nations to reduce their emissions. In so doing, it tried to address legitimate questions about equity and fairness. Poor nations argued — justifiably — that wealthy countries have profited immensely from fossil fuels, and that they were responsible for the bulk of historical greenhouse-gas emissions. They asserted their right to focus on lifting people out of poverty, while wealthy countries concentrated on bringing emissions down and developing technologies to enable everybody else to follow. It was a reasonable proposition — but it was destined to fail.

Emissions have continued to rise. Although most of the past emissions have come from wealthy nations, the bulk of those in the future will come from developing countries. Scientists have made it abundantly clear that every country must do everything that it can, and as fast as it can, if the world is to prevent the worst consequences of global warming.

The Paris agreement seeks to bridge the divide with carrots rather than sticks. Although countries agreed to engage in this new process, any action that they take to reduce emissions is on a purely voluntary basis. Indeed, the final change to the agreement in Paris, which took place quietly just minutes before the text was adopted, was to replace a ‘shall’ with a ‘should’ in a line stating how developed countries will commit to reducing emissions. This shift towards a voluntary framework based on national commitments was a necessary first step to bring everybody on board — and it worked.

Things may yet unravel. When negotiations pick up next year, the

first task will be to spell out exactly what information countries need to submit regarding their emissions and commitments, and how the review process will work. Given that there are no penalties for failing to achieve a commitment, the foundation of this agreement is transparency.

Governments, scientists and advocacy groups need solid information to verify that everybody is living up to their commitments and to transfer knowledge about what works and what doesn’t. The last — and often overlooked — piece of this puzzle is that developing countries will need help to establish the academic and technical expertise needed to meet these new international standards.

The Paris agreement represents a bet on technological innovation and human ingenuity. If governments follow through, companies and investors will shift resources towards clean energy to secure a place in an economy that will look very different several decades on.

In many ways, the debate about the long-term temperature-rise goal is symbolic. In the end, as noted in the agreement itself, the world needs to reduce net greenhouse-gas emissions to zero — and to do that, all countries must seek to halt the rise and bring down their emissions as soon as possible. Everybody has a role in making that happen. But today, the world can celebrate a win for global diplomacy. ■

**“The Paris agreement represents a bet on technological innovation and human ingenuity.”**

## Crop conundrum

*The EU should decide definitively whether gene-edited plants are covered by GM laws.*

When philosopher George Santayana said more than a century ago that those who do not learn from history are doomed to repeat it, he could have been predicting the European Union and its approach to genetically modified (GM) organisms.

As we report in a News story on page 319, the EU is dragging its feet over a legal ruling that could affect research and innovation for years to come. At stake is the use of gene-editing tools such as CRISPR–Cas9, which are revolutionizing biology. These techniques should theoretically trigger few safety alarms, yet they may be snared by the onerous legislation that has already added layers of bureaucracy to research involving conventional genetic engineering, and has slowed the cultivation of GM crops almost to a standstill in many nations.

The new tools can be applied to create mutations that could have occurred naturally, and leave no trace of foreign genes in the product.



Accordingly, the US Department of Agriculture has ruled in several cases that the products do not have to be regulated as GM organisms.

The European Commission is yet to send the same signal. In fact, it could decide that such products are governed by the existing cumbersome rules — its 2001 directive on the deliberate release of GM organisms into the environment. That would be a disaster for research.

The commission represents the interests of 28 member states, which are deeply divided on issues of genetic modification. But it needs to make clear — soon and with no room for misinterpretation — that work with these new techniques is important and does not necessarily need to be regulated in the same way as the previous generation of GM crops.

The precise and efficient gene-editing tools insert a gene that can create tiny, targeted mutations in an organism's own genome. These mutations can permanently change the function of a host gene, change its sensitivity to environmental cues or switch it off entirely; the foreign gene can then be bred out.

The core legal issue is whether the 2001 directive applies to all products of genetic engineering, or only to organisms that have been altered in a way that could not occur naturally. Clauses in the directive mention both.

Non-governmental organizations that are hostile to genetic engineering say that the directive is about the process by which products are created. But legal analyses conducted in the past year by several member states — including Germany, which has been opposed to conventional GM crops — concluded that it is fundamentally about the products themselves.

The commission's own legal analysis, being handled behind firmly closed doors, is the one that will count. But the result has been repeatedly delayed, spreading immense uncertainty in the scientific community.

It is now promised before the end of March. Why is it taking so long?

The commission has strongly hinted that the matter will ultimately be settled in court. Its decision, when it comes, is bound to annoy parties on one side, which may then sue. The possibility that a decision that releases many gene-edited products from GM regulation could be overturned by a court will add to the community's uncertainty.

There is some history here, and it should not be repeated. The commission tried, and failed, to resolve the lengthy disagreement over conventional GM crops by getting the European Court of Justice to rule on whether member states should be required to allow cultivation of such crops deemed safe by EU regulatory authorities. The court ruled that they should, but some countries banned it anyway. In a messy compromise, the EU now allows individual states to opt out.

The commission may be calculating that the reaction to a court ruling could be different this time, as a result of member states signalling their willingness to consider gene-edited products to be non-GM.

But letting a court decide a political issue is a poor option. It could take years. Even a positive verdict could rebound by reinforcing the narrative in some countries that the technology is being forced upon them. And it does not convey a positive message about legislation, which is supposed to reflect the will of the people.

The commission should indicate that the spirit of the 2001 directive does not cover the impact of the new gene-editing tools, and should give them an appropriate green light — with encouraging enthusiasm. If the exact wording of the 2001 directive gives room for doubt, then it should be updated to reflect a world in which new science has long overtaken the old.

Whatever the decision, the uncertainty must be lifted to allow research to proceed, and quickly. ■

# Science for peace

*The German research community can benefit from the influx of migrants.*

**T**his year's refugee crisis — a result of the civil war in Syria and enduring instability in the Middle East and Africa — has become an acid test for the European Union.

Although some countries would rather pull up the drawbridge where refugees are concerned, Germany has generously welcomed nearly one million migrants this year, without regard for the costs or logistical burden involved. "We can do it!" Chancellor Angela Merkel never failed to remind German citizens.

However, as police, immigration authorities, communities and volunteers creak under the strain, Merkel's optimism is increasingly being denounced in some quarters. To integrate hundreds of thousands of traumatized, mostly Muslim, war refugees into Western society is a massive social challenge. But, contrary to what some critics seem to assume, early signs show that the young refugees — and under-25s make up around half of the influx — will not be inclined to accept social welfare and sit back idly for long. Robbed of their hopes and dreams at home, many will grasp the opportunities offered.

And many will be eager to learn. If admitted into Germany's well-oiled education and science system (and into its booming labour market at large), they can be a boon rather than a burden to the country's knowledge-based economy.

German universities and science organizations are aware of the responsibility to these displaced people and the opportunity they represent. The messages they send in favour of openness and pluralism — defining features of any honest science — are laudable at a time

when xenophobia is on the rise elsewhere.

Thanks to several programmes and initiatives launched by the German science community in recent months, refugee students can access university education and doctoral-research opportunities, and qualified refugee scientists and scholars can participate in advanced science at research institutes across Germany (see page 320). These initiatives are much-needed and deserve every respect.

Refugees are expected to continue to arrive in Europe in large numbers, often lacking documentation of their professional or academic qualifications. Opportunities must continue to be available to them, and more must be helped to connect with potential employers, in and outside of academia.

Online tools such as the European Commission's Science4Refugees portal, on which employers can post job opportunities and refugees seeking science jobs can put their CVs, are well meant but not (yet) frequently used. Learned academies, universities and science organizations throughout Europe should more clearly and proactively promote the message that students, scholars and scientists who have been forced to flee their home can rebuild their careers as well as their lives.

Social researchers who study education, mobility and integration — for whom the current wave of migration is a research opportunity — must strive to empirically challenge presumptions about refugees' allegedly low level of qualification and susceptibility to political or religious extremism. To be sure, these things need to be — and will be — thoroughly investigated. But the idea touted by some that Muslim values are a fundamental obstacle to successful integration into a modern secular society is wrong and hopelessly short-sighted.

Whatever critics might say, Germany's rebirth as a haven for the persecuted is a powerful gesture of peace. Embracing refugees, while assuring anxious citizens that openness need not threaten their own quality of life, is perhaps the most pressing social challenge faced by science in these times. ■

➔ **NATURE.COM**  
To comment online,  
click on Editorials at:  
[go.nature.com/xhunqv](http://go.nature.com/xhunqv)

ANDREA ARMSTRONG



## Use data to challenge mental-health stigma

Web surveys of attitudes towards mental illness reveal the size of the problem — and offer a way to find fixes, says Neil Seeman.

The US National Institute of Mental Health considers stigma to be the most debilitating aspect of a mental illness. It is easy to see why. Stigma increases mental distress and leads to shame, avoidance of treatment, social isolation, and, consequently, a deterioration in health.

What form does this stigma take? Is it decreasing for mental illnesses such as depression, as claimed by some media articles? How can it be combated? We don't know the answers to those questions. That is partly because not enough people have asked them — and partly because not enough people have answered them. Surveys are expensive, and funds, especially for research on mental illness, are limited.

Surveys in the old days saw pollsters with hand-held clipboards quizzing shoppers in department stores. This gave way to the ubiquitous telephone survey. Today, the Internet affords ever more ways to collect survey data. Some years ago, I developed a way to ask questions in an efficient and global manner. It is called Random Domain Intercept Technology and it relies on people — like you — making mistakes while browsing the Internet. Mistyped URLs and broken web links trigger the survey, and invite the user to participate.

Unlike surveys in which people are given cash or rewards to answer questions, this method does not allow for a long-form questionnaire, although it can break down long surveys into shorter mini-surveys. It permits brief questions — often 8 to 15 of them — to be asked, and answered on a voluntary, non-incentivized basis by large numbers of random and anonymous people using the Internet. And that means almost everywhere in the world.

From September 2013 until May this year, we used the technology to ask some simple questions about mental illness and stigma. More than 1 million people from 229 territories responded. Their responses offer a unique and real-time snapshot of how the globe thinks about the estimated one-quarter of its population who will experience mental ill health (N. Seeman *et al.* *J. Affect. Disord.* **190**, 115–121; 2016).

The survey requested age and gender, and then asked two specific questions. First, is there someone you interact with every day who suffers from mental illness? (This may include psychosis, depression or addiction.) And second, are people who suffer from mental illness any of the following: more lazy, more violent, suffering from a condition as serious as physical illness, the victims of bad parenting, or able to overcome their challenges through 'tough love'?

In developed countries, only 7% of respondents thought that people with mental illness were more violent than the general population. In remarkable contrast, about 15% of those in developing countries thought that people with mental illness were more violent. Although 45–51% of respondents from developed countries believed that mental illness is similar to physical illness, only 7% of the same people thought that mental illness can be overcome. It seems that the understanding that mental illness has a biological cause makes the public more, rather than less, pessimistic about outcome. This has been reported previously, and is, at first glance, counterintuitive. Attributing illness to genes takes away blame, but at the same time, takes away hope for change.

Although the identity of individual respondents is unknown, the overall reproducibility of responses from any one region is high. When the same questions were posed every month in India for 21 months running, 10% of respondents each time reported that people with mental illness are more violent than others.

And despite the fact that mental illness is often a taboo subject, the anonymity of the survey facilitated consistent answers. In China, for example, people with mental illness are often viewed as bringing shame on their family. The 'loss of face' associated with mental illness there and in many developing countries attaches not only to the ill person, but also to family members. In this context it makes sense, therefore, that people with mental illness are kept at home, and this may explain the high proportion of people in China who reported having daily contact with a mentally ill person.

The approach I describe can uncover views on any topic held by those in Internet-enabled areas, currently 43% of the planet. And it can allow for 'before and after' surveys, assessing the effectiveness of population-wide interventions.

For instance, it would be of immense value to repeat this stigma survey in a region that has introduced a public-education anti-stigma campaign. The tool is not limited to stigma — in the field of mental health, for instance, it can probe suicidal ideas and, again, evaluate a suicide-prevention intervention. It can probe symptoms of post-traumatic stress disorder in the wake of a disaster (such as a hurricane or the Paris terrorist attacks) and test ways to mitigate these traumas.

Measuring a social problem on the scale of mental-illness stigma does not make it go away. But at least it shows us the size of the challenge — and could very well help to find ways to fix it. ■

Neil Seeman is chief executive of the RIWI Corporation and a senior fellow at Massey College, University of Toronto, Toronto, Canada.  
e-mail: [neil@riwi.com](mailto:neil@riwi.com)

THE  
ANONYMITY  
OF THE  
SURVEY  
FACILITATED  
CONSISTENT  
ANSWERS.

➔ [NATURE.COM](http://NATURE.COM)  
Discuss this article  
online at:  
[go.nature.com/lsux4f](http://go.nature.com/lsux4f)



# RESEARCH HIGHLIGHTS

Selections from the  
scientific literature

## ASTROPHYSICS

### Cosmic boost reveals dim galaxy

Astronomers have spied the faintest object ever seen in the early Universe.

Leopoldo Infante at the Pontifical Catholic University of Chile in Santiago and his team used NASA's Hubble and Spitzer space telescopes to study distant objects. They examined sections of the sky through a dense cluster of galaxies, which bends and magnifies incoming light, and found 22 faint galaxies. The oldest one was observed as it was 13.4 billion years ago, around 400 million years after the Big Bang.

The small, dim galaxy was named Tayna, meaning 'firstborn' in the Native South American language Aymara. It may be more representative of the first galaxies than other distant, brighter examples, say the authors.

*Astrophys. J.* 815, 18 (2015)

## GENOMICS

### Missed mutations in cancer genomes

A comparison of cancer-genome sequences produced by 18 different research teams reveals that less than half of cancer-linked mutations were identified by all the groups. This suggests that differences in experimental procedures and analysis could reduce the accuracy of cancer-genome sequencing, which is increasingly used in the clinic.

Ivo Gut at Spain's National Centre for Genomic Analysis in Barcelona, together with researchers in the International Cancer Genome Consortium, looked for genetic differences in cancerous and healthy tissue from the same person. They then compared these results

with a benchmark that used ten times more sequencing data than usual. Out of more than 1,200 single-letter mutations, only 40% were identified by all 18 teams.

DNA preparation and other parameters can be optimized easily to improve sequencing accuracy, the authors say.

*Nature Commun.* 6, 10001 (2015)

## GEOPHYSICS

### Rising sea levels alter Earth's spin

Researchers have confirmed that rising sea levels caused by melting glaciers are

slowing Earth's rotation.

As ice melts, it redistributes mass across the planet's surface, slightly changing the rate at which Earth spins. But a 2002 study could not explain the observed rotational changes on the basis of its assumptions about rising sea levels. Now Jerry Mitrovica of Harvard University in Cambridge, Massachusetts, and his colleagues say that they have resolved the problem. They used updated numbers for global sea-level rise, which are lower than those assumed in the 2002 study, and recalculated how the geographic poles have shifted

over the past 3,000 years.

The work improves scientists' understanding of how Earth's rotation has changed in the past, and how rising sea levels might continue to alter it in the future.

*Sci. Adv.* 1, e1500679 (2015)

## ASTRONOMY

### Galaxies caught in cosmic web

Astronomers have discovered eight massive young galaxies within what might be a large web of dark matter.

Ordinary matter, including



RICHARD WHITCOMBE/ALAMY

## ENVIRONMENTAL SCIENCE

### Ocean plastic piling up fast

Up to 240,000 tonnes of plastic particles are polluting the world's oceans — at least three times more than previous estimates.

Each year, 5 million to 13 million tonnes of plastic ends up in the sea, where it slowly degrades into microplastic particles that threaten marine ecosystems. Erik van Sebille at Imperial College London and his colleagues analysed 40 years of data on plastic collected from surface-trawling

plankton nets — more information than in previous studies. By combining those data with sophisticated ocean-circulation models, they estimated that the oceans contain 93,000–236,000 tonnes of microplastic particles.

This represents just 1% of ocean plastic: the rest lies intact (pictured) on the sea floor or shore, or trapped in marine organisms, the authors suggest.

*Environ. Res. Lett.* 10, 124006 (2015)

galaxies, is thought to have aggregated along threads of dark matter in the early Universe. But the progenitors of today's galaxies are often shrouded in clouds of dust, making it difficult for astronomers to spot them and test this theory.

Hideki Umehata at the European Southern Observatory in Garching, Germany, and his colleagues used the high-resolution Atacama Large Millimeter/submillimeter Array in Chile to make detailed observations of a narrow slice of the sky. They compared their results with previous surveys of the region to find the galaxies, which were more than 3.4 billion parsecs (11 billion light years) away and producing hundreds of millions of new stars each year. The study supports the idea that big galaxies form in areas with a high concentration of dark matter.

*Astrophys. J. Lett.* 815, L8 (2015)

## OCEAN SCIENCE

## Possible pause in Arctic sea-ice loss

An expected slowdown of large-scale heat circulation in the Atlantic Ocean could temporarily halt the decline of Arctic sea ice (**pictured**).

Stephen Yeager at the National Center for Atmospheric Research in Boulder, Colorado, and his colleagues used an Earth-system model to analyse the causes of decadal trends in sea-ice extent in the North Atlantic. They found that the drastic retreat of sea ice since 1990 coincided with a strong Atlantic circulation

that brought warm surface water from the tropics to high latitudes. If this circulation were to weaken, as observations suggest that it will, less heat arriving in the Arctic Ocean will probably lead to a pause in winter sea-ice loss over the next 5 to 10 years, the authors conclude.

They add, however, that the rate of sea-ice melting could jump back up afterwards as global warming continues. *Geophys. Res. Lett.* <http://doi.org/9wz> (2015)

## EVOLUTION

## How birds spread around the globe

The common ancestor of all modern birds lived in South America some 95 million years ago.

Birds inhabit every continent, and are among the most diverse vertebrate groups on Earth. To chart birds' rise and spread, Santiago Claramunt and Joel Cracraft at the American Museum of Natural History in New York created an evolutionary tree based on DNA sequences from 230 bird species and fossil records for 130 extinct species.

They found that bird diversity expanded rapidly after the demise of dinosaurs some 66 million years ago, dispersing along two primary routes. From South America, birds moved into North America, spread to Eurasia through the Arctic and then on to Africa. Birds arrived in Australia by way of Antarctica. *Sci. Adv.* 1, e1501005 (2015)

## MATERIALS

## Electrons dance in pulled graphene

Stretching an atom-thick strip of carbon could mimic the effects of a magnetic field, changing the behaviour of electrons so that the effect is 100 times stronger than that from normal magnets.

Teng Li at the University of Maryland in College Park and his colleagues calculated

## SOCIAL SELECTION

Popular topics on social media

## Deleting journal names triggers debate

Michael Eisen has long argued that research papers should be judged on the basis of their content, not on which journal they were published in. On 6 December, Eisen — a biologist at the University of California, Berkeley, and co-founder of the open-access publisher PLOS — decided to prove his point. He revamped his laboratory's website and announced on Twitter: "Made a new lab website — completely scrubbed any mention of journal titles — <http://www.eisenlab.org/publications.html>." A few other scientists followed suit, and one even went a step further. Plant geneticist Jeffrey Ross-Ibarra at the University of California, Davis, tweeted: "Following @mbeisen, removed journal names from website. But also links to cites, almetrics, [sic] & preprints. <http://www.rilab.org/pubs.html>." Others were sceptical. Manolis Dermitzakis, a geneticist at the University of Geneva, Switzerland, posted: "I don't see the point. The paper is published in a journal so this is just artificial. Or publish your papers on your website only."

➔ NATURE.COM

For more on popular papers: [go.nature.com/uwjpkf](http://go.nature.com/uwjpkf)

how to engineer the large pseudomagnetic fields that are produced when graphene is pulled from two ends. This strains bonds between carbon atoms, causing their electrons to move in a way that is similar to what happens in a magnetic field. The team found that a small tug (of up to 15% stretch) on certain shapes of graphene strip could produce a strong, nearly uniform field.

The designer shapes could help researchers to study the properties of graphene under extreme conditions — such as large magnetic fields — that are usually unattainable, the authors say.

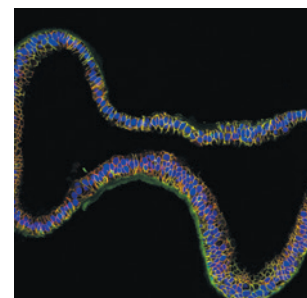
*Phys. Rev. Lett.* 115, 245501 (2015)

## DEVELOPMENTAL BIOLOGY

## Mini Fallopian tubes in a dish

Human Fallopian tubes contain adult stem cells that, when grown in the lab, can form miniature 3D structures resembling Fallopian tubes (**pictured**).

Thomas Meyer at the Max Planck Institute for Infection Biology in Berlin and his



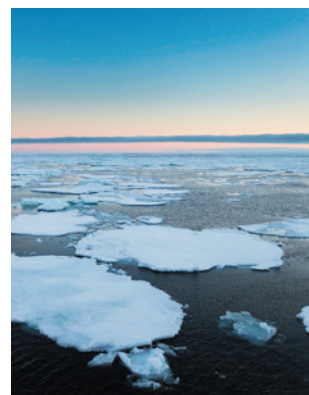
colleagues isolated cells from human Fallopian-tube samples and grew them in 3D cultures. Two weeks later, they saw mature 'organoids' that had folds in the tissue, hair-like structures called cilia, and secretory cells — all characteristics of the Fallopian tube. The organoids were stable for more than 16 months and sensitive to the hormones oestradiol and progesterone.

The organoids could be used to study tube pathology and certain types of ovarian cancer that are thought to originate in the Fallopian tubes, the authors say.

*Nature Commun.* <http://doi.org/9wr> (2015)

➔ NATURE.COM

For the latest research published by Nature visit: [www.nature.com/latestresearch](http://www.nature.com/latestresearch)





# SEVEN DAYS

The news in brief

## POLICY

### EU data-mining

The European Commission confirmed on 9 December that it wants to propose legislation to exempt certain types of text and data mining from copyright laws. As part of wider copyright reform, public-interest research organizations would be allowed to mine text and data from journal articles for research purposes without having to ask permission from the copyright owner. Researchers worried about legal restrictions on the data mining have long campaigned for the change.

### Gain of function

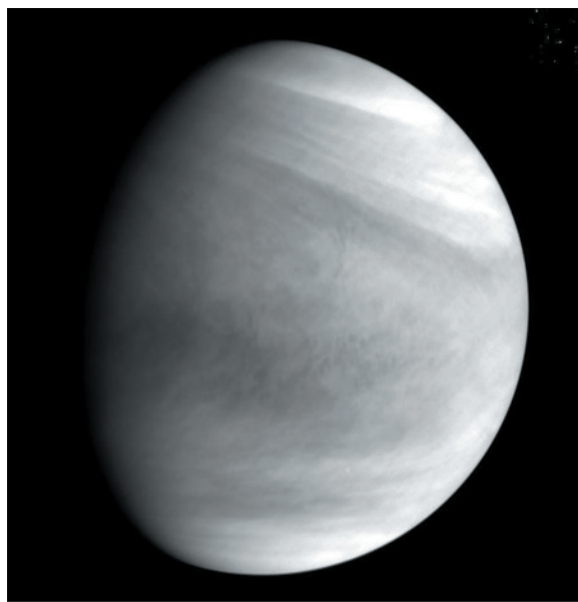
The US National Science Advisory Board for Biosecurity will convene on 7 January in Bethesda, Maryland, to assess the risks and benefits of 'gain-of-function' research — work intended to increase the virulence, transmissibility or host range of pathogens. The meeting will consider the findings of a 1,006-page risk-benefit assessment by the Gryphon Scientific consultancy in Takoma Park, Maryland, published on 11 December. The United States introduced a moratorium in October last

## NUMBER CRUNCH

# 51 trillion

The upper estimate on how many pieces of plastic smaller than 5 millimetres across had accumulated in the world's oceans by 2014. The lower estimate is 15 trillion.

Source: E. van Sebille *et al.* *Environ. Res. Lett.* **10**, 124006 (2015).



## Venus probe enters orbit

Japan's Akatsuki probe is circling Venus on an even-closer orbit than mission managers had hoped for, the Japan Aerospace Exploration Agency announced on 9 December. In 2010, Akatsuki missed its first chance to enter into orbit; it made a second, successful attempt this month. At its closest approach, the probe will fly just 400 kilometres above Venus's surface, from which point researchers aim to study the planet's atmosphere. Three of the craft's five cameras have already been confirmed as functional after their extra five years in space. This image was taken by the ultraviolet imager from about 72,000 kilometres above Venus's surface.

year on federal funding of such research on the agents that cause influenza, severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS).

## EVENTS

### Paris deal done

Negotiations at the Paris climate-change talks sealed a deal between 195 nations to limit warming to "well below" 2 °C above pre-industrial temperatures. The 32-page package was made on 12 December after 2 weeks of talks, and

commits most nations to significant reductions in carbon emissions. The agreement notes that vulnerable low-lying countries are set to face rising sea levels and stronger storms. See page 315 for more.

### Open-data accord

Four international science lobby groups have launched a joint accord supporting open data as a tool for more-equitable science. The initiative, announced on 9 December in Pretoria during the first Science Forum

South Africa, attempts to make it easier for developing countries to participate in research on a global level. It is also the first attempt to unify the fragmented activities of the four bodies, which represent different disciplines and global regions: the International Council for Science, the InterAcademy Partnership, the International Social Science Council and the World Academy of Sciences.

## PEOPLE

### DOE science chief

Cherry Murray, a physicist at Harvard University in Cambridge, Massachusetts, will be the new director of the US Department of Energy (DOE) science office. The US Senate confirmed her appointment on 10 December. The decision is considered surprising because most recent federal appointments have been blocked by the Senate — the previous nominee for the office, Michael Kastner, was not confirmed after his 2013 nomination. Murray, an expert in condensed-matter physics and photonics, will take office this month.

## FUNDING

### AIDS funding cut

In a readjustment of priorities announced on 11 December, the US National Institutes of Health (NIH) will no longer put 10% of its science budget towards AIDS research, overturning a requirement of more than 20 years. The policy has been controversial, with opponents arguing that the number of HIV/AIDS deaths dropped precipitously during this time. The NIH director's advisory council said that, as existing grants end, the move will eventually free up hundreds of millions of dollars

JAXA

for research on other diseases. The agency will refocus its remaining AIDS budget away from basic biology and towards the creation of specific therapies and vaccines.

## Animal names safe

Thanks to gifts totalling \$1.35 million (US\$959,000), the International Commission on Zoological Nomenclature (ICZN) secretariat will be able to continue its role of ensuring that animal species are named in a systematic fashion. The commission had been facing insolvency. Based at the National University of Singapore, the ICZN enforces a globally accepted nomenclature code to ensure that each species has a unique and scientifically appropriate name; around 15,000 new species are described annually. The philanthropic Lee Foundation in Singapore provided nearly all of the endowment, the ICZN announced on 14 December in Berlin at a joint meeting with the International Union of Biological Sciences.

### FACILITIES

## Stellarator is go

The world's largest 'stellarator' fusion device roared into life on 10 December. The €1-billion (US\$1.1-billion) Wendelstein 7-X, based at the Max Planck



Institute for Plasma Physics in Greifswald, Germany, produced its first plasma (pictured), lasting for one-tenth of a second and reaching a temperature of around 1 million °C. Although the test run used helium, next year the device will start superheating hydrogen in experiments designed to explore the suitability of the technique for commercial fusion. The stellarator confines ionized gas using intricately interwoven magnetic coils. The design is difficult to construct but potentially a more stable alternative to the doughnut-shaped 'tokamak' used by the international ITER fusion project, based in southern France.

### BUSINESS

## NEON Inc. out

The US National Science Foundation (NSF) has decided to replace the manager of the beleaguered

US\$434 million National Ecological Observatory Network, the company NEON, Inc. The decision comes after the company told the NSF in June that it was running \$80 million over budget. That triggered a congressional hearing and warning from NSF that it might oust NEON, Inc. in favour of another operator. The construction of the remaining observatory sites will probably be overseen by another company.

## Chemicals combine

Two of the world's largest chemical and agricultural companies, Dow Chemical of Midland, Michigan, and DuPont of Wilmington, Delaware, will attempt to merge. On 11 December, the companies announced that subject to regulatory approval, they would combine forces to create a firm valued at US\$130 billion. That would then break apart into three independent companies: one focused on agriculture, another on materials science and the third on specialty products.

## Dengue vaccine

The first vaccine for preventing the tropical disease dengue fever has been approved for use in Mexico. The vaccine,

## COMING UP

### 18–21 DECEMBER

The European Society for Medical Oncology holds its Asia Congress in Singapore.

[go.nature.com/6vwgoh](http://go.nature.com/6vwgoh)

### 19–22 DECEMBER

The International Liposome Society gathers its members at University College London to discuss the use of liposomes in drug and vaccine delivery.

[go.nature.com/jmjGuy](http://go.nature.com/jmjGuy)

Dengvaxia, developed by Sanofi Pasteur of Lyon, France, was approved on 9 December by Mexico for patients aged 9 to 45 who live in areas where dengue is endemic. The viral infection is carried by mosquitoes, and the number of infections worldwide has risen rapidly in recent years. The vaccine protects against the four variants of the dengue virus, and was approved after a clinical-development programme that involved more than 40,000 people in 15 countries.

## Open intelligence

A group of individuals and companies from Silicon Valley in California have formed a non-profit company to research artificial intelligence (AI) that is "likely to benefit humanity as a whole". The company, OpenAI, has raised US\$1 billion and is co-chaired by Elon Musk, chief executive of the electric-car company Tesla Motors and private space-flight firm SpaceX. Musk has previously urged caution when it comes to AI, warning that it could become "more dangerous than nukes".

➔ [NATURE.COM](http://NATURE.COM)

For daily news updates see:

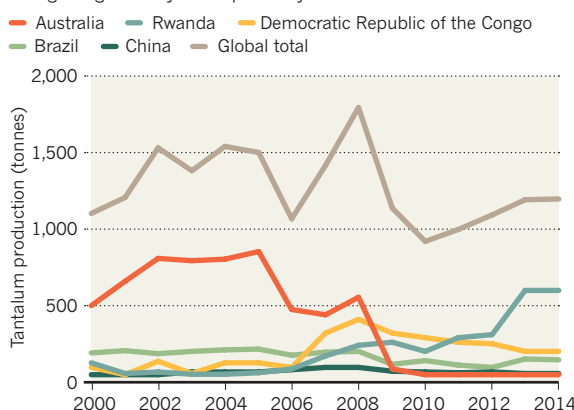
[www.nature.com/news](http://www.nature.com/news)

## TREND WATCH

The source of tantalum, a metal used in the electronics industry and for specialized mechanical parts, has shifted dramatically since 2000, according to a US Geological Survey report. In 2000, Australia was the world's main source of tantalum (producing 45%), but in 2014 Rwanda produced most (50%). Tantalum is a 'conflict mineral', meaning that its sale may finance conflict in countries such as the Democratic Republic of the Congo, and buyers must check the metal's source. See [go.nature.com/wog3zu](http://go.nature.com/wog3zu) for more.

### TANTALUM SOURCES SHIFT

The location of the biggest tantalum producers has changed significantly in the past 15 years.





# NEWS IN FOCUS

**CANCER** Roots of malignancy drive debate over role of 'bad luck' in disease **p.317**

**MARINE SCIENCE** China and South Korea agree to territorial talks **p.318**

**REFUGEE CRISIS** German social scientists help tap into immigrants' potential **p.320**



**COMMUNICATION** The scientific myths that evidence can't kill **p.322**

FRANÇOIS MORI/AP



French foreign minister Laurent Fabius, chairman of the Paris talks, gives the new climate accord two thumbs up.

## CLIMATE CHANGE

# Nations adopt historic global climate accord

*Agreement commits world to holding warming 'well below' 2 °C.*

BY JEFF TOLLEFSON & KENNETH R. WEISS, PARIS

When the gavel came down for the final time at the climate summit in Paris on 12 December, representatives from 195 countries erupted into cheers.

They had approved a landmark plan to combat climate change after two weeks of gruelling negotiations. The agreement commits most countries to reduce their

greenhouse-gas emissions, while seeking to protect low-lying islands from rising seas and helping poor nations to develop their economies without relying on cheap, dirty fossil fuels.

The accord, years in the making, seeks to hold warming "well below" 2 °C above pre-industrial temperatures. Countries' current climate pledges fall short of that goal, but many scientists and governments see the Paris agreement as the last, best hope to set the planet on a course to avoid catastrophic climate change.

"History is written by those who commit, not those who calculate," French president François Hollande told negotiators after the accord was adopted. "Today you have committed."

The ambitious 32-page package contains a multitude of provisions to accelerate the world's transition from fossil fuels to solar, wind, nuclear, hydropower and other clean energy sources.

Nearly every country is asked to play ►

► its part in ensuring that greenhouse-gas emissions peak, and then begin to decline, as soon as possible. Countries will assess their progress towards reducing emissions in 2018, and must revisit their climate pledges every five years, beginning in 2020. The aim is that these pledges will become more ambitious over time.

To ensure that countries are keeping to their commitments, the agreement creates a transparent system for measuring, reporting and verifying emissions, while allowing some flexibility for countries that have little capacity to do so. The plan allows for an independent technical review, and all but the smallest, poorest countries will have to report their emissions every two years. But negotiators have left many of the details to be debated at the next major climate talks, in 2016.

“On transparency, the agreement is a little bit loosey-goosey,” says Michael Oppenheimer, a climate scientist at Princeton University in New Jersey. “It could be turned into something that is very effective, but the delegates kicked the can down the road.”

Others worry about how developing countries can be helped to build their capacity to monitor emissions. “Transparency and governance are not something you obtain with a decree,” says Joseph Armathé Amougou, director of Cameroon’s National Observatory on Climate Change. He will be responsible for developing and reporting his country’s greenhouse-gas inventory, but he currently has neither the budget nor the employees to do so.

The Paris agreement includes non-binding language that outlines a plan for wealthy nations to increase their climate aid to poorer nations beyond their current commitment of US\$100 billion per year by 2020. And developing nations pushed successfully for the pact to recognize that vulnerable countries will face damage from rising seas, raging storms and other impacts of climate change.

The official recognition of damage was a huge achievement, says Mohamed Adow of Christian Aid, an advocacy group based in London. “We now have loss and damage as an integral part of the climate regime.” But the pact explicitly bars poorer nations from seeking compensation or from holding wealthy, major polluters liable for these losses.

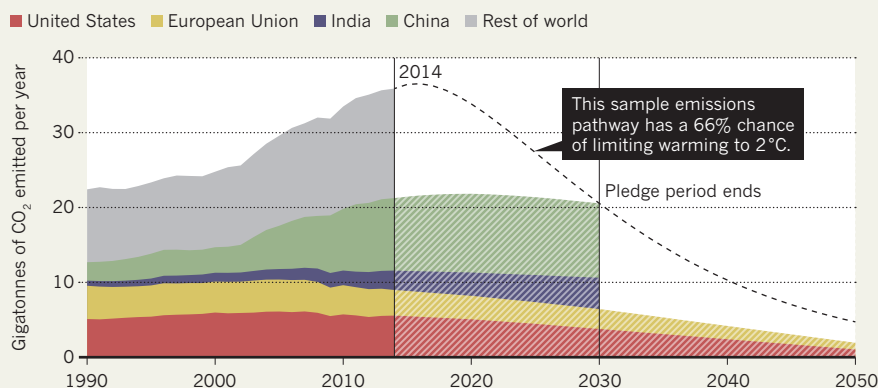
## AN ARDUOUS JOURNEY

Wearied negotiators, running on nervous energy and caffeine, approved the Paris agreement a day after their self-imposed deadline — and only after a major push by leaders of the United Nations and the host country.

In a soaring speech, Hollande implored delegates to pass an accord that would send “a message of life” to rebuke the perpetrators of the terrorist attacks that killed 130 people in Paris on 13 November. “I will be delighted, relieved, proud, that it be launched from Paris, because Paris was attacked almost exactly a month ago,” he said. “France asks you,

## TIGHT BUDGET

Major greenhouse-gas emitters have pledged to reduce their carbon footprints, but holding warming to 2°C will be a challenge.



SOURCE: GLOBAL CARBON PROJECT

calls upon you, to adopt the first universal agreement on climate.”

The long road to the Paris agreement began in Rio de Janeiro in 1992, when nations approved a general ‘framework’ to combat climate change that left the details for later agreements. After 20 annual meetings with little progress to curb ever-soaring emissions, representatives arrived in Paris with pledges from 187 countries that outlined the steps each would take to cut its emissions by 2030.

Never before had so many promises been on the table — but many pledges

were hedged with conditions, such as calls for financial aid to build alternative energy plants, save remaining forests or relocate people living in harm’s way. Even if all of the promises were fulfilled, and were followed by substantial additional emissions reductions, the world would warm 2.7°C by 2100 (see ‘Tight budget’). This is deep into the territory that scientists expect would prompt catastrophic, irreversible climate changes.

Yet the Paris agreement seeks to limit planetary warming to well below 2°C, urging nations to pursue an even stricter target, 1.5°C. To put this in perspective, the average global temperature has already risen 1°C since the start of the Industrial Revolution.

Many environmentalists say that the agreement and the goals are strong enough to create momentum and put pressure on governments moving forward. “We see the key elements that we have always said we need for a good agreement,” says Nathaniel Keohane, who heads the global climate programme for the Environmental

Defense Fund in New York City.

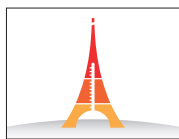
Others say that the Paris accord should prod businesses to pursue clean energy and green growth. “Markets now have the clear signal that they need to unleash the full force of human ingenuity and scale up investments that will generate low-emissions, resilient growth,” said United Nations secretary-general Ban Ki-moon. “What was once unthinkable has now become unstoppable.”

Climate scientists who gathered in Paris to observe the negotiations were pleased with the accord’s ultimate goal, but wanted more details about how nations would achieve significant emissions reductions. “This does not send a clear signal about the level and timing of emission cuts, and does not provide a useful yardstick against which to measure progress,” says Steffen Kallbekken, research director at the Center for International Climate and Energy Policy in Oslo. Although the Paris plan is not inconsistent with the science, he says, it does not reflect the best available research.

The Intergovernmental Panel on Climate Change (IPCC) has concluded that holding warming to 2°C will probably require emissions to be cut by 40–70% by 2050 compared with 2010 levels, Kallbekken notes. Achieving the 1.5°C target would require substantially larger emissions cuts — of the order of 70–95% by 2050.

The Paris agreement directs the IPCC to study scenarios for limiting warming to 1.5°C, and to deliver a report to nations by 2018 to help them determine how much to strengthen their climate commitments.

The fact that the accord prominently mentions the 1.5°C target is a huge victory for vulnerable countries, says Saleemul Huq, director of the International Centre for Climate Change and Development in Dhaka, Bangladesh. “Coming into Paris, we had all of the rich countries and all of the big developing countries not on our side,” says Huq, an adviser to a coalition of least-developed nations. “In the 14 days that we were here, we managed to get all of them on our side.” ■ SEE EDITORIAL P.307



➔ **NATURE.COM**  
For Nature’s full coverage of the Paris talks, see: [go.nature.com/c7146j](http://go.nature.com/c7146j)



## MEDICINE

# Cancer studies clash

*Researchers debate relative importance of environmental and intrinsic factors in malignancy development.*

BY HEIDI LEDFORD

Most cases of cancer result from avoidable factors such as toxic chemicals and radiation, contends a study published online in *Nature* on 16 December (S. Wu *et al.* *Nature* <http://dx.doi.org/10.1038/nature16166>; 2015). The paper attempts to rebut an argument that arose early this year, when a report in *Science* concluded that differences in inherent cellular processes are the chief reason that some tissues become cancerous more frequently than others (C. Tomasetti and B. Vogelstein *Science* **347**, 78–81; 2015).

The work led to assertions that certain forms of cancer are mainly the result of “bad luck”, and suggested that these types would be relatively resistant to prevention efforts. “There’s no question what’s at stake here,” says John Potter of the Fred Hutchinson Cancer Research Center in Seattle, Washington, who studies causes of cancer. “This informs whether or not we expend energy on prevention.”

In their *Science* paper, mathematician Cristian Tomasetti and cancer researcher Bert Vogelstein at Johns Hopkins University in Baltimore, Maryland, calculated the relationship between the number of stem-cell divisions and the risk of developing cancer in various tissues. Every instance of cell division comes with a risk that DNA will be incorrectly copied, leading to mutations — some of which could contribute to cancer. The duo’s analysis found a correlation: the more stem-cell divisions that occur in a given tissue over a lifetime, the more likely it is to become cancerous.

Tomasetti and Vogelstein then sorted types of cancer according to how much of the variability in risk is due to stem-cell divisions versus to some ‘extrinsic’ factor, such as environmental exposure to carcinogens. The authors argued that although some cancers clearly had strong environmental links — such as liver cancers caused by hepatitis C infection or lung cancer resulting from smoking — there were others for which the variation was explained mainly by defects in stem-cell division. In those cases, they argued, early detection and treatment would be more effective than prevention.

Something about that did not sit right with Yusuf Hannun, a cancer researcher at Stony Brook University in New York. “What they did was interesting, but I was startled by the conclusion,” he says.

The original work, Hannun and his colleagues argued, assumed that the two variables

— intrinsic stem-cell division rates and extrinsic factors — were entirely independent. But what if environmental exposures affect stem-cell division rates, as radiation is known to do?

## A DIFFERENT TAKE

Hannun and his team also used other lines of evidence to try to pinpoint the contribution of environmental factors to cancer risk. They looked at epidemiological data showing that, for example, people who migrate from regions of lower cancer risk to those with higher risk soon develop disease at rates consistent with their new homes. The authors also examined patterns in the mutations associated with certain cancers; ultraviolet light, for example, tends to create a tell-tale signature of mutations in DNA. And they used other mathematical models, expanding the data set used in the earlier work to include prostate and breast cancer — two of the most common cancers.

The models suggested that mutations during cell division rarely build up to the point of producing cancer, even in tissues with relatively high rates of cell division. In almost all cases, the team found that some exposure to

**“There’s no question what’s at stake. This informs whether or not we expend energy on prevention.”**

carcinogens or other environmental factors would be needed to trigger disease.

Tomasetti counters that he never intended to explain why cancers develop.

His analysis, he says, was based on normal stem-cell division in healthy tissue and was meant to explain only why some cancers are more prevalent than others. He also argues that the models created by Hannun and his colleagues make too many assumptions and fail to incorporate some features of tumour growth.

Some specialists in cancer prevention welcome the *Nature* paper because of fears that the public — and possibly also funders of scientific research — might conclude that prevention efforts are not worthwhile, says Edward Giovannucci, who studies cancer prevention at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts. “By not smoking, your lifetime risk of lung adenocarcinoma drops dramatically,” he says. “The fact that your risk of pelvic sarcoma is even lower because there’s less stem-cell division — so what?” ■



DONG-A ILBO/AFP/GETTY

Chinese fishing boats are pursued by a South Korean coastguard vessel (top right) in the Yellow Sea in 2011; the nations have overlapping claims in the region.

#### OCEANOGRAPHY

# Talks lift hopes in territorial impasse

*Negotiations between South Korea and China to demarcate Yellow Sea boundary could aid marine science.*

BY MARK ZASTROW, SEOUL

China and South Korea have scheduled talks for 22 December to address a decades-long boundary dispute that has hampered research and exploration in the Yellow Sea. This northern part of the East China Sea, between mainland China and the Korean peninsula, is home to a rich ecosystem that is under intense environmental strain from human activities.

Confrontations over fishing rights in the disputed region have turned deadly — and research is not immune to the tension. South Korean scientists report that the Chinese coastguard has intercepted research vessels in the Yellow Sea and East China Sea on at least ten occasions, threatening their activities and forcing them to move east. At other times, the Chinese navy has shadowed South Korean research vessels. “The confrontations are happening all the time,” says marine sedimentologist Kyung-Sik Choi of Seoul National University.

The friction in the Yellow Sea is one of many marine territorial disputes in east Asia: over

the past two years, China has captured the world’s attention with its construction of artificial islands in the South China Sea and a series of alleged ramblings of local fishing boats by its coastguard and navy vessels. A spat with Japan over islands and gas fields in the East China Sea is also escalating, as China boosts its military presence and extraction efforts there.

In this particular case, both parties seem ready — at least publicly — to seek a solution. Chinese President Xi Jinping and South Korean President Park Geun-hye pledged in July 2014 to begin talks by the end of 2015.

“If the maritime boundary is fixed in some way, it will be good for scientists because we will know exactly where our playground is,” says Hyun-Chul Han, a marine geologist at the Korea Institute of Geoscience and Mineral Resources in Daejeon. “It will be a great relief and secure scientists’ safety.”

Few expect South Korea and China to fully resolve their dispute in this first round of talks. But some analysts say that boosting scientific ties between the nations in the Yellow Sea would be a feasible — and politically valuable — initial step.

“Maybe this could be an area of low-hanging fruit that these talks could address, to at least point to some level of utility and productivity,” says James Schoff, a senior associate at the Carnegie Endowment for International Peace in Washington DC.

#### THE LAW OF THE SEA

Under the 1982 United Nations Convention on the Law of the Sea, nations can claim exclusive rights to exploit resources in an exclusive economic zone (EEZ) within 200 nautical miles (370 kilometres) of their coasts. But because the Yellow Sea is less than 400 nautical miles in breadth, China and South Korea’s EEZs overlap, and they have never agreed to a boundary (see ‘Troubled waters’). Research vessels from both countries avoid straying across a line of longitude about halfway between Seoul and Qingdao, effectively dividing the Chinese and South Korean marine-science communities. The law does not in principle restrict purely scientific activities in another nation’s EEZ, but in practice, countries can quickly set these zones off-limits to others.

Chinese data covering the Yellow Sea look “cut in half” because of the dispute, says Zuosheng Yang, a marine geologist at the Ocean University of China in Qingdao.

In the past, China has rejected simply drawing a line that is equidistant from the two nations’ coasts. Instead, it claimed rights to about two-thirds of the Yellow Sea, based on the extent to which sediments billowing out from China’s Huang He and Yangtze rivers blanket the sea floor. This ‘silt line’ was met with howls of protest from South Korean scholars and received little international support. But the silt line has a practical significance: Chinese boats motor across it to escape the turgid, fish-poor



sediment plumes, sometimes leading to fatal clashes with South Korea's coastguard. In 2011, a Chinese fisherman stabbed a Korean coastguard to death with a shard of broken window glass; in a separate 2014 skirmish, the Korean coastguard shot and killed a Chinese fisherman.

The dispute has also prevented cooperation in assessing the deterioration of the Yellow Sea's marine ecosystem. Dams in Chinese rivers have interrupted the once-steady flow of sediment and nutrients into the waters, and pollution has created enormous algal blooms. Urbanization has also claimed most of the tidal flats that once ringed the Yellow Sea basin, threatening key habitats for migratory birds.

Monitoring and management of the basin requires collaboration, says Paul Liu, an oceanographer at North Carolina State University in Raleigh. South Korean and Chinese ocean researchers do share some data through a joint marine-research centre in Qingdao, which has held workshops and coordinated some work since 1995. But when asked about the boundary dispute, Wei Zheng, the centre's vice-director, said: "It still is a problem." She declined to comment further, citing the sensitivity of the issue.

Choi, for example, says that he and his colleagues would like to conduct a deep seismic



survey transecting the entire Yellow Sea. But he says that the project would need permission and protection from China's coastguard to prevent passing fishing boats causing any damage to the kilometres-long cables and attached equipment.

Both Liu and Yang say that an agreement would similarly foster collaborations to look

at how sediments have swirled across the Yellow Sea in the past, and how new dams on China's rivers have changed that process. "The Chinese cannot only study the western side, or Koreans cannot only study the eastern side," Liu says. "They have to work together to know the whole picture of the area." ■

## POLICY

# Europe's genetically edited plants stuck in legal limbo

*Scientists frustrated at delay in deciding if GM regulations apply to precision gene editing.*

BY ALISON ABBOTT

Plant geneticist Stefan Jansson is championing at the bit to start field trials on crops tweaked with powerful gene-editing technologies. He plans to begin by using edits to study how the cress plant *Arabidopsis* protects its photosynthetic machinery from damage in excessively bright light.

But the future of his work depends on the European Commission's answer to a legal conundrum. Should it regulate a gene-edited plant that has no foreign DNA as a genetically modified (GM) organism?

Jansson, who works at Umeå University in Sweden, says that he will drop his experiments if the plants are classed as GM, because Europe's onerous regulations would make his work too expensive and slow. He and many others are anxiously awaiting the commission's decision, which will dictate how they approach experiments using the latest gene-editing techniques,

including the popular CRISPR-Cas9 method.

The commission has repeatedly stalled on delivering its verdict, which will apply to edited animals and microorganisms as well as plants. It now says that it will make its legal analysis public by the end of March. Swedish authorities, meanwhile, have told Jansson that unless the commission specifies otherwise, they will not require his cress to be subject to GM regulations.

## GENETIC EDITING

The legal limbo is having a big impact on research, says René Smulders of the plant-breeding division at Wageningen University and Research Centre in the Netherlands. He says that this year, he was rejected for a European Union grant — on changing the composition of a plant's oils by editing a gene — because referees were concerned about the legal uncertainty. "Some scientists hesitate to start using the new methods in case they end up being regulated and their research

projects hit a dead end," he says.

At issue is the interpretation of a 2001 European Commission directive on releasing GM organisms into the environment, which covers field trials and cultivation. It defines GM organisms as having alterations that cannot occur naturally, which were made by genetic engineering.

What is unclear is how this relates to experiments, such as Jansson's, in which researchers introduce foreign DNA to direct a precise edit in a plant's own genetic material but then use selective breeding to remove the foreign gene. The final plant has a few tweaked nucleotides, but cannot be distinguished from a wild plant that might have acquired the same mutation naturally — so it cannot be traced in the environment as EU regulations require.

Many EU member states — including Sweden — have conducted their own analyses of the directive, and argue that it should not apply to edited plants that do not contain foreign DNA. But some non-governmental organizations ►

► (NGOs) hostile to genetic manipulation have produced analyses that conclude the directive should apply because genetic engineering is involved.

Academic scientists and seed and crop companies fear that plants made with the latest gene-editing techniques may share the fate of conventional GM plants in Europe. Strict regulations, cumbersome bureaucracy and activism against GM organisms have meant that scientists in some countries, such as Germany, do not even attempt field trials. The regulations have increased the costs of bringing a GM crop to market, and many European nations do not allow such crops to be cultivated at all. That is frustrating for plant scientists who want their work to be useful to the world, says Jonathan Jones, a plant researcher at the Sainsbury Laboratory in Norwich, UK.

"We hoped that the new plant-breeding techniques would offer ways of achieving the same outcome without the onerous regulations — and fear that might not turn out to be the case," he says.

Many countries outside Europe do not face the same uncertainty, because they regulate GM organisms according to the nature of the product, not how it was made. In the United States, gene-edited crops containing no foreign genetic material are assessed on a case-by-case basis. In 2004, the biotechnology company Cibus, based in San Diego, California, was told that the US Department of Agriculture would not need to regulate its herbicide-resistant oilseed rape, made with an earlier form of gene-editing. Its crop is now cultivated in the United States. (The White House did, however, begin a review of all US biotechnology regulation in July.)

Since 2011, Cibus has asked six countries — Finland, Germany, Ireland, Spain, Sweden and the United Kingdom — whether they would consider its crop to come under the scope of the EU directive. Without guidelines from the commission, each conducted its own analysis and said that it would not. Cibus has now done field trials in the United Kingdom and Sweden, but it put its activities on hold after the commission sent a letter to all EU member states on 15 June, asking them to wait for its legal interpretation.

Whatever the commission decides, it is likely that either a member state, an NGO or a company will sue — meaning that the European Court of Justice may make the final, binding decision on the matter.

Many plant scientists do basic research, so their gene-edited plants never need to leave the greenhouse. But Jansson must plant his cress outside to test its photosynthetic abilities in natural conditions. With his country's approval, he plans to plant the crop in the spring. "Lawyers talk and talk — I think it is important for Europe to have a test case," he says. ■ [SEE EDITORIAL P.307](#)

## EUROPE

# German researchers pledge refugee help

*Social scientists launch integration studies and warn of need to counter rising xenophobia.*

BY QUIRIN SCHIERMEIER

After civil war broke out in Syria, Mohammad Khamis lost his parents and his home — but not his dream of becoming a scientist. In July 2013, he boarded a flight from Damascus, where he had studied electrical engineering, to Egypt. In Alexandria, he paid traffickers about €5,000 (US\$5,500) for a boat passage to Europe. The 9-day voyage to the Italian island of Lampedusa, on an unseaworthy sloop with 100 other desperate refugees, was a nightmare of fear, vomit and thirst.

Two years later, Khamis, now 22, is attending classes in maths, physics and chemistry at the Technical University of Munich (TUM) in Germany, where he sought asylum in August 2013 and was last year accepted as a war refugee. "There is no future for me in Syria," he says on a cold December day in Munich. "I would like to stay here to study and find a good research job. My dream is to discover something new."

Social scientists studying the flow of refugees into Germany want to discover something themselves: how many of the incoming people are, like

Khamis, well-qualified, motivated and eager to learn — a boon for the economy. These migration researchers say that Germany has become a case study in the difficulties of suddenly integrating a large group of culturally diverse foreigners into a society; the nation has registered nearly one million asylum-seekers this year, more than half of them from Syria. It is the highest such influx in Western Europe.

After a short-lived wave of hospitality in September, when chancellor Angela Merkel promised that Germany would be a welcoming host to the persecuted, many citizens and some right-leaning politicians have begun to voice concerns, painting a picture of a Muslim-dominated parallel society of poorly trained recipients of social welfare.

Research may be able to counter the rising tide of xenophobia and aid the urgent process of resettling refugees by revealing more about migrants' skills and cultural values, says

David Schiefer, a Berlin-based psychologist with a German advisory body on migration and integration who is planning interviews with refugees. "We need to give these people a voice," he says.

With about half of the newcomers under 25 years of age, Germany's higher-education and science systems have a particular obligation — and the well-funded capacity — to help, say researchers. "Science has a responsibility to help tackle the huge integration challenge ahead," says Alexander Kurz, head of human resources at the Fraunhofer Society in Munich, which runs centres for applied research. "There is great readiness among our staff of 25,000 scientists from 100 nations to provide mentorship and practical help."

## LISTENING TO REFUGEES

Reliable data on refugees' qualifications and backgrounds are lacking. "We're poking around in the fog," says Ludger Wößmann, a director of the Ifo Center for the Economics of Education in Munich. International assessments of 15-year-olds suggest that up to two-thirds of Syrian refugee students might lack basic reading, writing and maths skills, he says. German industrial groups say that the large majority of migrants have minimal skills and poor language abilities, making them hardly employable.

But these assumptions are ill-informed, says Steven Vertovec, director of the Max Planck Institute for the Study of Religious and Ethnic Diversity in Göttingen. In fact, the newcomers are probably as diverse as German society at large, he says. "There are many highly educated, secularized people among the Syrians, Iraqis and Afghans who are seeking asylum here."

Vertovec is leading a study in Lower Saxony in northern Germany that aims to interview asylum-seekers to examine their needs and aspirations, as well as to uncover best practices for responding to refugees. The goal is to produce practical guidelines for city workers and volunteer social workers in asylum-seeker camps on how to work with groups of migrants who may differ enormously in age, religion, language and education status. "Successful integration requires a nuanced understanding of migrants' backgrounds and values," he says.

**"Science has a responsibility to help tackle the huge integration challenge ahead."**

Students such as Khamis (who officially has 'guest' status at TUM; he is not yet formally enrolled in Germany's university system) are not an uncommon sight in the country's university lecture halls. TUM has about 100 guest students; across the country, there are a few thousand. To help universities to cope with the influx, the government in November approved an extra €100 million for student counselling, language training and stipends.

### GOVERNMENT SUPPORT

On 11 December, Germany's main research-funding agency, the DFG, encouraged grant holders to consider hiring refugee scientists in their research. DFG-funded scientists whose work would benefit from the participation of qualified academics or PhD students among the refugees are free to submit supplemental proposals for 'guest funding', said DFG president Peter Strohschneider.

In a strategy paper seen by *Nature*, a group from seven Max Planck institutes, in response to a call for research ideas by the society's president, Martin Stratmann, has outlined a variety of research needs around humanitarian migration, from international law and human-rights issues to health and gender studies.

Marie-Claire Foblets, director of the Max Planck Institute for Social Anthropology in Halle, plans to ask a culturally diverse group



UJI BENZ/TUM

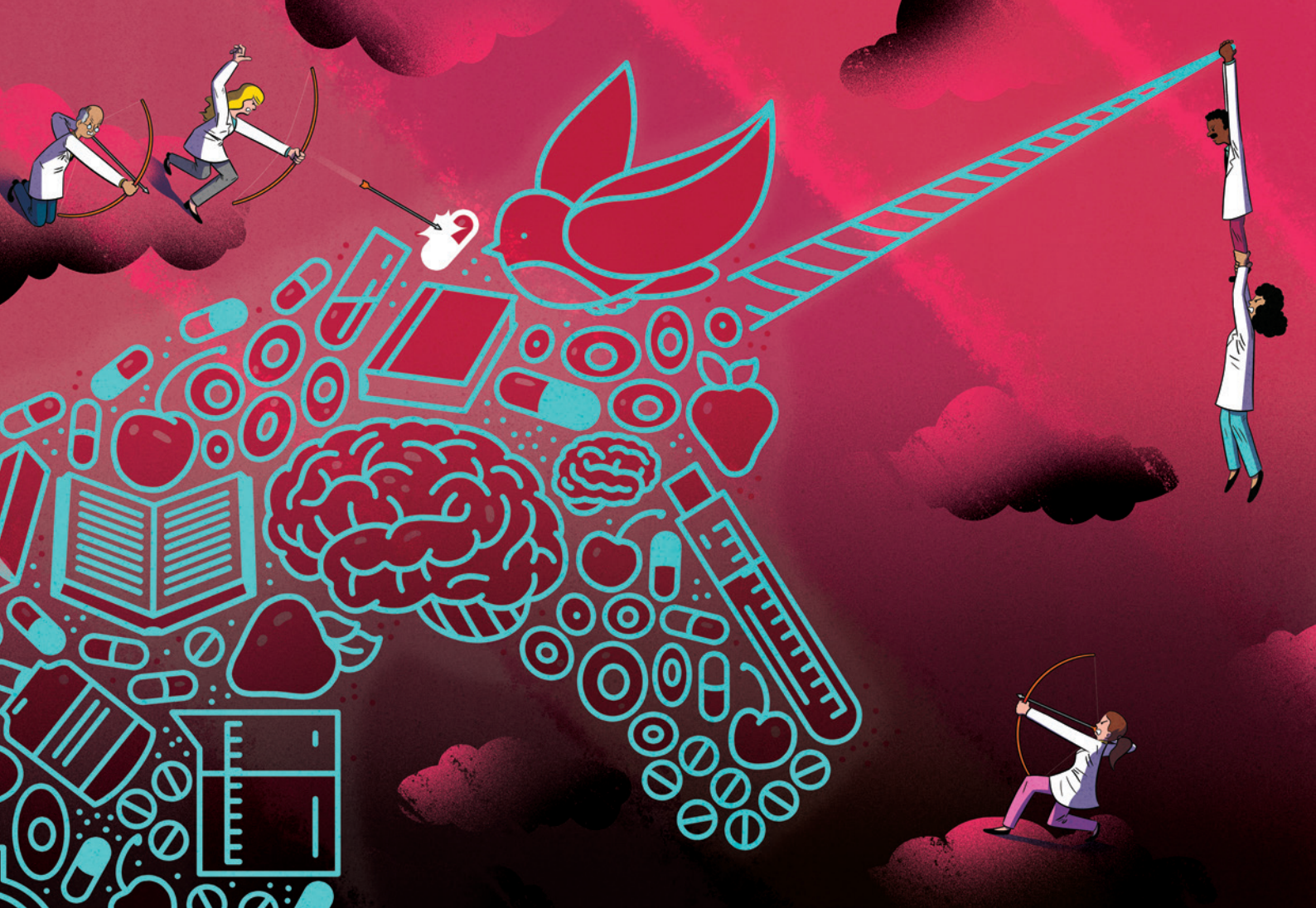
Mohammad Khamis (centre), who left Syria in 2013, is now attending the Technical University of Munich.

of refugees — including guest students at the University of Halle-Wittenberg — for accounts of their lives and experiences. Other questions, such as those concerning refugees' citizenship and civil rights, the potential lure of extremism, and the fate of children who might be staying with radicalized parents, will require the involvement of law experts, criminologists,

educators and others, she says.

Khamis, for one, is happy to write up the story of his life for research. Having passed German-language tests, he hopes to enrol at the university next term as a regular student. "Germany has been good to me," he says. "Now that my life can start again I do hope that I can give something back." ■ **SEE EDITORIAL P.308**





# Myths *that will not die*

False beliefs and wishful thinking about the human experience are common. They are hurting people — and holding back science.

BY MEGAN SCUDELLARI

**I**n 1997, physicians in southwest Korea began to offer ultrasound screening for early detection of thyroid cancer. News of the programme spread, and soon physicians around the region began to offer the service. Eventually it went nationwide, piggybacking on a government initiative to screen for other cancers. Hundreds of thousands took the test for just US\$30–50.

Across the country, detection of thyroid cancer soared, from 5 cases per 100,000 people in 1999 to 70 per 100,000 in 2011. Two-thirds

of those diagnosed had their thyroid glands removed and were placed on lifelong drug regimens, both of which carry risks.

Such a costly and extensive public-health programme might be expected to save lives. But this one did not. Thyroid cancer is now the most common type of cancer diagnosed in South Korea, but the number of people who die from it has remained exactly the same — about 1 per 100,000. Even when some physicians in Korea realized this, and suggested that thyroid screening be stopped in 2014, the Korean

ILLUSTRATIONS BY RYAN SNOOK

Thyroid Association, a professional society of endocrinologists and thyroid surgeons, argued that screening and treatment were basic human rights.

In Korea, as elsewhere, the idea that the early detection of any cancer saves lives had become an unshakeable belief.

This blind faith in cancer screening is an example of how ideas about human biology and behaviour can persist among people — including scientists — even though the scientific evidence shows the concepts to be false. “Scientists think they’re too objective to believe in something as folklore-ish as a myth,” says Nicholas Spitzer, director of the Kavli Institute for Brain and Mind at the University of California, San Diego. Yet they do.

These myths often blossom from a seed of a fact — early detection does save lives for some cancers — and thrive on human desires or anxieties, such as a fear of death. But they can do harm by, for instance, driving people to pursue unnecessary treatment or spend money on unproven products. They can also derail or forestall promising research by distracting scientists or monopolizing funding. And dispelling them is tricky.

Scientists should work to discredit myths, but they also have a responsibility to try to prevent new ones from arising, says Paul Howard-Jones, who studies neuroscience and education at the University of Bristol, UK. “We need to look deeper to understand how they come about in the first place and why they’re so prevalent and persistent.”

Some dangerous myths get plenty of air time: vaccines cause autism, HIV doesn’t cause AIDS. But many others swirl about, too, harming people, sucking up money, muddying the scientific enterprise — or simply getting on scientists’ nerves. Here, *Nature* looks at the origins and repercussions of five myths that refuse to die.

### MYTH 1: SCREENING SAVES LIVES FOR ALL TYPES OF CANCER

Regular screening might be beneficial for some groups at risk of certain cancers, such as lung, cervical and colon, but this isn’t the case for all tests. Still, some patients and clinicians defend the ineffective ones fiercely.

The belief that early detection saves lives originated in the early twentieth century, when doctors realized that they got the best outcomes when tumours were identified and treated just after the onset of symptoms. The next logical leap was to assume that the earlier a tumour was found, the better the chance of survival. “We’ve all been taught, since we were at our mother’s knee, the way to deal with cancer is to find it early and cut it out,” says Otis Brawley, chief medical officer for the American Cancer Society.

But evidence from large randomized trials for cancers such as thyroid, prostate and breast has shown that early screening is not the

lifesaver it is often advertised as. For example, a Cochrane review of five randomized controlled clinical trials totalling 341,342 participants found that screening did not significantly decrease deaths due to prostate cancer<sup>1</sup>.

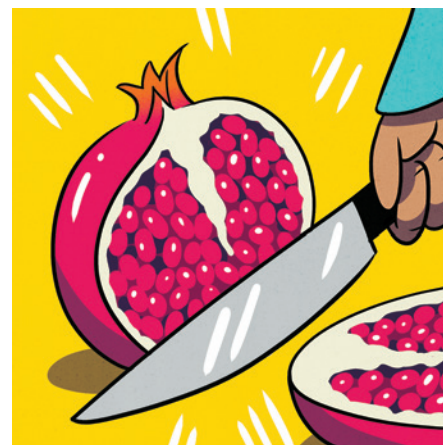
“People seem to imagine the mere fact that you found a cancer so-called early must be a benefit. But that isn’t so at all,” says Anthony Miller at the University of Toronto in Canada. Miller headed the Canadian National Breast Screening Study, a 25-year study of 89,835 women aged 40–59 years old<sup>2</sup> that found that annual mammograms did not reduce mortality from breast cancer. That’s because some tumours will lead to death irrespective of when they are detected and treated. Meanwhile, aggressive early screening has a slew of negative health effects. Many cancers grow slowly and will do no harm if left alone, so people end up having unnecessary thyroidectomies, mastectomies and prostatectomies. So on a population level, the benefits (lives saved) do not outweigh the risks (lives lost or interrupted by unnecessary treatment).

*“We cherry-pick the numbers that put us on top.”*

Still, individuals who have had a cancer detected and then removed are likely to feel that their life was saved, and these personal experiences help to keep the misconception alive. And oncologists routinely debate what ages and other risk factors would benefit from regular screening.

Focusing so much attention on the current screening tests comes at a cost for cancer research, says Brawley. “In breast cancer, we’ve spent so much time arguing about age 40 versus age 50 and not about the fact that we need a better test,” such as one that could detect fast-growing rather than slow-growing tumours. And existing diagnostics should be rigorously tested to prove that they actually save lives, says epidemiologist John Ioannidis of the Stanford Prevention Research Center in California, who this year reported that very few screening tests for 19 major diseases actually reduced mortality<sup>3</sup>.

Changing behaviours will be tough. Gilbert Welch at the Dartmouth Institute for Health Policy and Clinical Practice in Lebanon, New Hampshire, says that individuals would rather be told to get a quick test every few years than be told to eat well and exercise to prevent cancer. “Screening has become an easy way for both doctor and patient to think they are doing something good for their health, but their risk of cancer hasn’t changed at all.”



### MYTH 2: ANTIOXIDANTS ARE GOOD AND FREE RADICALS ARE BAD

In December 1945, chemist Denham Harman’s wife suggested that he read an article in *Ladies’ Home Journal* entitled ‘Tomorrow You May Be Younger’. It sparked his interest in ageing, and years later, as a research associate at the University of California, Berkeley, Harman had a thought “out of the blue”, as he later recalled. Ageing, he proposed, is caused by free radicals, reactive molecules that build up in the body as by-products of metabolism and lead to cellular damage.

Scientists rallied around the free-radical theory of ageing, including the corollary that antioxidants, molecules that neutralize free radicals, are good for human health. By the 1990s, many people were taking antioxidant supplements, such as vitamin C and  $\beta$ -carotene. It is “one of the few scientific theories to have reached the public: gravity, relativity and that free radicals cause ageing, so one needs to have antioxidants”, says Siegfried Hekimi, a biologist at McGill University in Montreal, Canada.

Yet in the early 2000s, scientists trying to build on the theory encountered bewildering results: mice genetically engineered to overproduce free radicals lived just as long as normal mice<sup>4</sup>, and those engineered to overproduce antioxidants didn’t live any longer than normal<sup>5</sup>. It was the first of an onslaught of negative data, which initially proved difficult to publish. The free-radical theory “was like some sort of creature we were trying to kill. We kept firing bullets into it, and it just wouldn’t die,” says David Gems at University College London, who started to publish his own negative results in 2003 (ref. 6). Then, one study in humans<sup>7</sup> showed that antioxidant supplements prevent the health-promoting effects of exercise, and another associated them with higher mortality<sup>8</sup>.

None of those results has slowed the global antioxidant market, which ranges from food and beverages to livestock feed additives. It is projected to grow from US\$2.1 billion in 2013 to \$3.1 billion in 2020. “It’s a massive racket,”



says Gems. “The reason the notion of oxidation and ageing hangs around is because it is perpetuated by people making money out of it.”

Today, most researchers working on ageing agree that free radicals can cause cellular damage, but that this seems to be a normal part of the body’s reaction to stress. Still, the field has wasted time and resources as a result. And the idea still holds back publications on possible benefits of free radicals, says Michael Ristow, a metabolism researcher at the Swiss Federal Institute of Technology in Zurich, Switzerland. “There is a significant body of evidence sitting in drawers and hard drives that supports this concept, but people aren’t putting it out,” he says. “It’s still a major problem.”

Some researchers also question the broader assumption that molecular damage of any kind causes ageing. “There’s a question mark about whether really the whole thing should be chucked out,” says Gems. The trouble, he says, is that “people don’t know where to go now”.

### MYTH 3: HUMANS HAVE EXCEPTIONALLY LARGE BRAINS

The human brain — with its remarkable cognition — is often considered to be the pinnacle of brain evolution. That dominance is often attributed to the brain’s exceptionally large size in comparison to the body, as well as its density of neurons and supporting cells, called glia.

None of that, however, is true. “We cherry-pick the numbers that put us on top,” says Lori Marino, a neuroscientist at Emory University in Atlanta, Georgia. Human brains are about seven times larger than one might expect relative to similarly sized animals. But mice and dolphins have about the same proportions,

## MYTHS THAT PERSIST

### Irksome misbeliefs

*Nature* polled doctors and scientists for the medical myths that they find most frustrating. Here’s what turned up.

#### Vaccines cause autism

Although there are some risks associated with vaccines, the connection to neurological disorders has been debunked many times over.

#### Paracetamol (acetaminophen) works through known mechanisms

Although it is widely used, there are only hints as to how it and other common drugs actually work.

#### The brain is walled off from the immune system

The brain has its own immune cells, and a lymphatic system that connects the brain to the body’s immune system has recently been discovered.

#### Homeopathy works.

It doesn’t.

and some birds have a larger ratio.

“Human brains respect the rules of scaling. We have a scaled-up primate brain,” says Chet Sherwood, a biological anthropologist at George Washington University in Washington DC. Even cell counts have been inflated: articles, reviews and textbooks often state that the human brain has 100 billion neurons.

More accurate measures suggest that the number is closer to 86 billion. That may sound like a rounding error, but 14 billion neurons is roughly the equivalent of two macaque brains.

Human brains are different from those of other primates in other ways: *Homo sapiens* evolved an expanded cerebral cortex — the part of the brain involved in functions such as thought and language — and unique changes in neural structure and function in other areas of the brain.

The myth that our brains are unique because of an exceptional number of neurons has done a disservice to neuroscience because other possible differences are rarely investigated, says Sherwood, pointing to the examples of energy metabolism, rates of brain-cell development and long-range connectivity of neurons. “These are all places where you can find human differences, and they seem to be relatively unconnected to total numbers of neurons,” he says.

The field is starting to explore these topics. Projects such as the US National Institutes of Health’s Human Connectome Project and the Swiss Federal Institute of Technology in Lausanne’s Blue Brain Project are now working to understand brain function through wiring patterns rather than size.

### MYTH 4: INDIVIDUALS LEARN BEST WHEN TAUGHT IN THEIR PREFERRED LEARNING STYLE

People attribute other mythical qualities to their unexceptionally large brains. One such myth is that individuals learn best when they are taught in the way they prefer to learn. A verbal learner, for example, supposedly learns best through oral instructions, whereas a visual learner absorbs information most effectively through graphics and other diagrams.

There are two truths at the core of this myth: many people have a preference for how they receive information, and evidence suggests that teachers achieve the best educational outcomes when they present information in multiple sensory modes. Couple that with people’s desire to learn and be considered unique, and conditions are ripe for myth-making.

“Learning styles has got it all going for it: a seed of fact, emotional biases and wishful thinking,” says Howard-Jones. Yet just like sugar, pornography and television, “what you prefer is not always good for you or right for you,” says Paul Kirschner, an educational psychologist at the Open University of the Netherlands.

In 2008, four cognitive neuroscientists reviewed the scientific evidence for and against learning styles. Only a few studies had rigorously put the ideas to the test and most of those that did showed that teaching in a person’s preferred style had no beneficial effect on his or her learning. “The contrast between the enormous popularity of the learning-styles approach within education and the lack of credible evidence for its





utility is, in our opinion, striking and disturbing,” the authors of one study wrote.

That hasn't stopped a lucrative industry from pumping out books and tests for some 71 proposed learning styles. Scientists, too, perpetuate the myth, citing learning styles in more than 360 papers during the past 5 years. “There are groups of researchers who still adhere to the idea, especially folks who developed questionnaires and surveys for categorizing people. They have a strong vested interest,” says Richard Mayer, an educational psychologist at the University of California, Santa Barbara.

In the past few decades, research into educational techniques has started to show that there are interventions that do improve learning, including getting students to summarize or explain concepts to themselves. And it seems almost all individuals, barring those with learning disabilities, learn best from a mixture of words and graphics, rather than either alone.

Yet the learning-styles myth makes it difficult to get these evidence-backed concepts into classrooms. When Howard-Jones speaks to teachers to dispel the learning-styles myth, for example, they often don't like to hear what he has to say. “They have disillusioned faces. Teachers invested hope, time and effort in these ideas,” he says. “After that, they lose interest in the idea that science can support learning and teaching.”

#### MYTH 5: THE HUMAN POPULATION IS GROWING EXPONENTIALLY (AND WE'RE DOOMED)

Fears about overpopulation began with Reverend Thomas Malthus in 1798, who predicted that unchecked exponential population growth would lead to famine and poverty.

But the human population has not and is not growing exponentially and is unlikely to do so, says Joel Cohen, a populations researcher at the Rockefeller University in New York City. The world's population is now growing at just half the rate it was before 1965. Today there are an estimated 7.3 billion people, and that is projected to reach 9.7 billion by 2050. Yet beliefs that the rate of population growth will lead to some doomsday scenario have been continually perpetuated. Celebrated physicist Albert Bartlett, for example, gave more than 1,742 lectures on exponential human population growth and the dire consequences starting in 1969.

The world's population also has enough to eat. According to the Food and Agriculture Organization of the United Nations, the rate of global food production outstrips the growth of the population. People grow enough calories in cereals alone to feed between 10 billion and 12 billion people. Yet hunger and malnutrition persist worldwide. This is because about 55% of the food grown is divided between feeding cattle, making fuel and other materials or going to waste, says Cohen. And what remains is not evenly distributed — the rich have



plenty, the poor have little. Likewise, water is not scarce on a global scale, even though 1.2 billion people live in areas where it is.

“Overpopulation is really not overpopulation. It's a question about poverty,” says Nicholas Eberstadt, a demographer at the American Enterprise Institute, a conservative think tank based in Washington DC. Yet instead of examining why poverty exists and how to sustainably support a growing population, he says, social scientists and biologists talk past each other, debating definitions and causes of overpopulation.

Cohen adds that “even people who know the facts use it as an excuse not to pay attention to the problems we have right now”, pointing to the example of economic systems that favour the wealthy.

Like others interviewed for this article, Cohen is less than optimistic about the chances of dispelling the idea of overpopulation and other ubiquitous myths (see ‘Myths that persist’), but he agrees that it is worthwhile to try to prevent future misconceptions. Many myths have emerged after one researcher extrapolated beyond the narrow conclusions of another's work, as was the case for free radicals. That “interpretation creep”, as Spitzer calls it, can lead to misconceptions that are hard to excise. To prevent that, “we can make sure an extrapolation is justified, that we're not going beyond the data”, suggests Spitzer. Beyond

that, it comes down to communication, says Howard-Jones. Scientists need to be effective at communicating ideas and get away from simple, boiled-down messages.

Because once a myth is here, it is often here to stay. Psychological studies suggest that the very act of attempting to dispel a myth leads to stronger attachment to it. In one experiment, exposure to pro-vaccination messages reduced parents' intention to vaccinate their children in the United States. In another, correcting misleading claims from politicians increased false beliefs among those who already held them. “Myths are almost impossible to eradicate,” says Kirschner. “The more you disprove it, often the more hard core it becomes.” ■

**Megan Scudellari** is a science journalist in Boston, Massachusetts.

1. Ilic, D., Neuberger, M. M., Djulbegovic, M. & Dahm, P. *Cochrane Database Syst Rev.* **1**, CD004720 (2013).
2. Miller, A. B. et al. *Br. Med. J.* **348**, g366 (2014).
3. Saquib, N., Saquib, J. & Ioannidis, J. P. A. *Int. J. Epidemiol.* **44**, 264–277 (2015).
4. Doonan, R. et al. *Genes Dev.* **22**, 3236–3241 (2008).
5. Pérez, V. I. et al. *Aging Cell* **8**, 73–75 (2009).
6. Keaney, M. & Gems, D. *Free Radic. Biol. Med.* **34**, 277–282 (2003).
7. Ristow, M. et al. *Proc. Natl Acad. Sci. USA* **106**, 8665–8670 (2009).
8. Bjelakovic, G., Nikolova, D. & Gluud, C. *J. Am. Med. Assoc.* **310**, 1178–1179 (2013).
9. Pashler, H., McDaniel, M., Rohrer, D. & Bjork, R. *Psychol. Sci. Publ. Int.* **9**, 105–119 (2008).



# COMMENT

**CITIES** Share time, land and skills for urban well-being **p.330**

**CHEMISTRY** An aesthetic and scientific celebration of the snowflake **p.331**

**COMMUNICATION** How to influence policy and still get tenure **p.332**



**METRICS** Equal first authors lose in citation processes **p.333**

ILLUSTRATIONS BY DAVID PARKINS



## Why synthesize?

**Philip Ball** ponders the many reasons that chemists make molecules, and weighs what is lost, and gained, when they don't.

**W**hy do chemists make molecules? The obvious (and true) answer is: because we need them. That is why chemical synthesis is still vibrant, and will continue to supply the drugs, materials and commodities of the twenty-first century. Every year brings its bounty. In 2015, chemists published a new and elegant route to the anticancer drug paclitaxel (Taxol)<sup>1</sup>, and syntheses of a nodulisporic acid that might act as an insecticide<sup>2</sup> and, in this journal, of an anti-HIV alkaloid<sup>3</sup>.

There are also less utilitarian reasons for making molecules. One chemist might want to explore theoretical questions, such as what constitutes a bond. Another might delight in, and be curious about, the variety of shapes and structures that molecules can have. That diversity of purpose is how it

should be. For at the root of the impulse to build molecules is a deep, cherished belief that arguably distinguishes chemistry from other sciences: that there is an art in making, worth nurturing for its own sake.

Chemical synthesis can entail many things — minor modification of existing molecular frameworks, for example, or making new materials. Total synthesis — the complete construction of a complex (often natural) molecule from simple reagents — has long been seen as the epitome of the art. But some say that the age of monumental projects to make complicated molecules is waning. These long and expensive procedures may produce tiny yields of the target molecule. And now there are automated methods that put molecules together; eventually, even the synthetic route might be

planned automatically.

So, could bespoke, elaborate synthesis become a boutique rarity akin to the hand-crafting of books in the age of e-readers and print-on-demand? And if synthesis is relegated to a routine, should chemists be worried?

Chemists periodically revisit (and revile) the argument over whether total synthesis is moribund, generally with more heat than light. It's the wrong argument. Both the methods and motives of chemistry are evolving fast. We should be focusing on how synthesis responds. That response may be driven partly by pragmatism. But synthesis also has pedagogical and — unusually in a core scientific discipline — aesthetic dimensions that must be factored into the equation. There are several possible reasons to make ►

► complex molecules by total synthesis. A century ago the aim was often to identify a molecular structure, as in Robert Robinson's classic work on the synthesis of strychnine in the 1940s: if you know what happens at each step, you know what the end result looks like. That motive has vanished, however, thanks to advances in structural analysis, especially crystallography and nuclear magnetic resonance spectroscopy.

Another reason that chemists synthesized natural products was because of their useful properties. Molecules could be cheaper to make from scratch than to extract painstakingly from rare organisms. The total synthesis of the dye indigo in the 1870s that led to the collapse of the cultivation of the indigo plant is a canonical historical example.

Today, most wholly synthetic routes to complex natural products are too complicated to be useful in themselves to the pharmaceutical industry. Even the celebrated total synthesis of paclitaxel in 1994 was never seriously expected to lead to a commercial route (it is now made semi-synthetically from a natural precursor, or by fermentation). But total synthesis of a natural product can give chemists access to non-natural derivatives that might have pharmacological effects — as, for example, in the discovery of new antibiotics.

What's more, the grounding in synthetic chemical methods provided by making a complex natural molecule from scratch is said to equip students with the practical skills that industry requires. Synthesis also cultivates an understanding of the basic principles of chemistry: how and why reactions occur, the relationships between molecular shape and function, and so on. An ability to synthesize molecules remains essential training for the next generation of chemists; it is simply part of the indispensable core of the subject. By the same token, a lack of drawing skill does not make an artist bad but it makes them limited.

Perhaps that's why chemists with synthesizing skills are often said to get jobs in the pharmaceutical industry most easily. What is less clear is whether these skills can be learnt only by tackling fiendishly complicated structures. Indeed, Derek Lowe at Vertex Pharmaceuticals in Boston, Massachusetts, argues that drug companies value not the synthetic prowess per se but the concomitant ability to solve problems fast — and to cope with the inevitable disappointments, because most drugs, like most organic reactions, do not work without a lot of tinkering.

George Whitesides at Harvard University in Cambridge, Massachusetts, raises a different concern. He worries that training US graduate students to do organic synthesis when most of it is now being done in China, risks equipping them for jobs that do not exist. In this view, molecule-building is just

another kind of manufacturing technology: if it can be done more cheaply elsewhere, it is best not even to try to compete, just to outsource.

In any case, the utility of resulting skills and products is only part of the argument advanced for why chemical synthesis matters. Great synthetic chemists of the mid-to-late twentieth century, such as Robert Woodward and Elias Corey, are revered not so much for what they made but for how they made it: for the way they refined the art. Woodward argued<sup>4</sup> that an innate aesthetic appeal is involved: "The unique challenge which chemical synthesis provides for the creative imagination and the skilled hands ensures that it will endure as long as men write books, paint pictures, and fashion things which are beautiful, or practical, or both."

These notions are part of the lore of the field. Milestones of synthesis are recounted in heroic terms, their pathways examined step by step as exemplars of elegant strategy. The comparison is often made with games of chess: victory is seen as a triumph of personal style and flair. One team of expert total synthesizers has more recently justified the pursuit by saying<sup>5</sup> that it "demands the following virtues from, and cultivates the best in, those who practice it: ingenuity, artistic taste, experimental skill, persistence, and character ... its dual nature as precise science and fine art provides excitement and rewards of rare heights". The baroque carbon frameworks that still grace the pages of chemistry journals are often presented with a virtuoso flourish.

### BUILD IT WELL

Nonetheless some chemists feel that total synthesis of large and complicated natural products has now become a scaling of peaks just because they are there — with, moreover, a meaningless race to the summit that is often won by brute force. Lowe calls this the "human-wave-attack style" of making gigantic natural products, which, he jokes, ends in papers reporting the total synthesis of a molecule that no one much cares about, "made in a way you'd figure would probably work, using reactions everyone already knows".

He contends that useful chemistry — a new method of making bonds, say — is rarely discovered along the way, partly because the field is so competitive. No one is going to dawdle to search for clever shortcuts if they can just follow tried and tested paths. When some enormous and intricate natural product becomes the next Everest, elegance is sacrificed for speed, and ingenuity for graduate-student hours, Lowe says.

Advocates of total synthesis retort that priority races and showboating — who can make the hardest molecule fastest — are less common now. The aim is no longer just to build the desired structure but to build it well. For example, chemists seek a route that is economical in atoms (producing few waste products and side reactions), environmentally friendly and sustainable. As Steven Ley of the University of Cambridge, UK, put it in 2007 after completing a 22-year effort to synthesize the complicated natural insecticide azadirachtin, "I don't have to be first; the elegance of the approach is what interests me" (see *Nature* **448**, 630–631; 2007).

Thanks to the efforts of the giants of synthesis past and present, almost any molecule can now be made in principle. The question is whether it can be made in a practical and fruitful way.

### COLLECTIVE COMPLEXITY

To some chemists then, making complex molecules for their own sake no longer seems the pinnacle of craft. That arguably reflects changes in the objectives of chemistry as a whole. Whitesides has suggested<sup>6</sup> that if chemistry is regarded as a science of atoms and individual molecules, then its low-hanging fruits are gone. The future of chemistry, according to him, lies with complex molecular systems that display collective properties and functions at a range of size scales. This may be the only means by which chemistry can fulfil its obligations in areas ranging from medicine to materials, energy and information.

Take the much bewailed drying-up of the drugs pipeline. Although the reasons are complicated, one factor could be that the old model of developing and refining a single drug molecule by a long process of screening and clinical trials is no longer the best option. The future of molecular medicine might instead include suites of molecules performing operations in concert, as biomolecules do in the cell. This, after all, is how the transformative gene-editing technique of CRISPR–Cas9 works.

Moreover, the complexity and versatility of life's molecules come not from a huge array of synthetic substrates and reactions, but from combinations of a rather small set of parts, assembled through a limited arsenal of bond-forming processes and guided by natural selection. Certainly, natural products of extreme intricacy can result. But theoretical and experimental surveys of 'chemical space' — the astronomical array of possible molecules — give no reason to think that ornate solutions are essential or unique.

Complicated natural products with synthetically challenging frameworks do not tend to feature in nature's methods of making or transforming energy, replicating, information processing, locomotion

*"Like architecture, chemistry deals in elegance in both design and execution."*



or much else. Work like that of David Liu and collaborators at Harvard<sup>7</sup> shows that nature's synthetic principles of information-guided templating coupled to variation and selection might be a productive way to make useful synthetic molecules. In fact, that approach has also yielded new ways of assembling them<sup>8</sup>: new bond-forming chemistry, which was found by explicitly looking for it and not by hoping that it would emerge in the course of scaling a molecular Eiger. Such work suggests that, even though molecule-building is sure to remain a crucial part of the chemical enterprise, conventional organic synthesis need not be the only, or even the best, way to do it.

### AUTOMATING THE ART

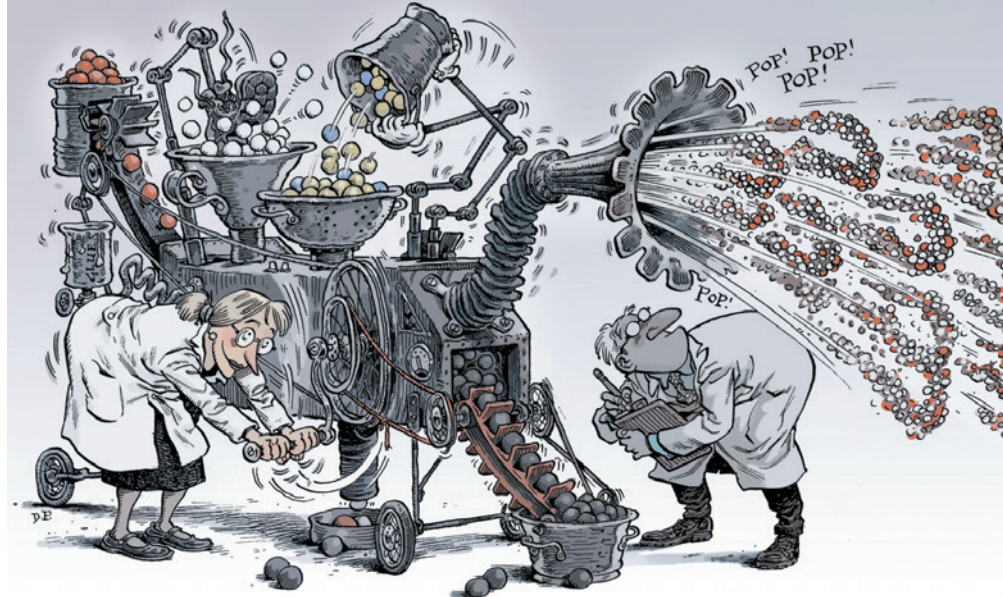
One of the common criticisms of total synthesis is that it rarely offers a route that the chemical or pharmaceutical industries can use: it takes too long, there are too many steps, the yields are too low and the costs too high. If you want to make a complicated molecule, do you really need an army of dedicated graduate students working through the night? Or could it be done by machine?

Automated synthesis is already possible for peptides and nucleic acids, which can be obtained by mail order with essentially any sequence. Oligosaccharides are also yielding to this approach. As a result we have lucrative peptide and oligonucleotide drugs, and glycoprotein drugs are on the way. Work<sup>9</sup> by Martin Burke at the University of Illinois at Urbana-Champaign suggests that a great variety of small and medium-sized organic molecules could be made this way too.

Burke uses a single, general-purpose reaction to assemble carbon-framework building blocks. He deploys the Suzuki coupling, in which a boronic acid substituent on one carbon reacts with a halogen substituent on the other in the presence of a palladium catalyst. The crucial trick is first to control this process for stepwise assembly<sup>10</sup>, and then to automate the procedure by trapping the products of each step on silica beads to extract and release them for the next step. It is not by any means possible to build everything this way. But the method gives access to an impressive array of molecules rapidly and cheaply at the push of a button. Burke and his colleagues have used it to make less toxic derivatives of the antifungal natural product amphotericin B.

Automation is nothing new. Microfluidic flow processes for conducting multistep syntheses without the need for purification at each step have been used for at least a decade. And with a small repertoire of standard, reliable bond-forming reactions, even the synthetic strategy itself could conceivably now be planned by machine.

The idea that synthesis could become the workaday cranking-out of any structure is



disturbing to anyone brought up to regard it as an art. It seems akin to the notion that artificial intelligence will one day compose our music and write our novels. But the 'art' of chess has been overtaken by brute-force number-crunching. There is no fundamental reason why chemical synthesis should be any different — nor, in fact, why machine-learning should not one day find superior, smoother and more efficient synthetic strategies than we can intuit (see *Nature* **512**, 20–22; 2014).

If that happens, some magic would be lost. But there could be practical gains. Today we need to make many molecules fast, to outpace the rise of antibiotic resistance, for example. This is acknowledged by the Dial-a-Molecule project, funded since 2010 by the UK Engineering and Physical Sciences Research Council, which aims to extend the assembly-line principle of oligonucleotide synthesis to any small organic molecule.

The project's vision is that "In 20–40 years, scientists will be able to deliver any desired molecule within a timeframe useful to the end-user, using safe, economically viable and sustainable processes" (see [www.dial-a-molecule.org](http://www.dial-a-molecule.org)). It aims to use computer algorithms to devise the best route for making a target molecule with a suite of 'click' reactions, which are efficient, predictable and dependable. The goal is to make any given molecule in a matter of days.

Easier synthesis could free chemists to think creatively about molecular design: to focus on the question of what is worth making. That is currently the other big obstacle to effective drug discovery. As Burke explains, we do not yet know the rules that nature uses to 'design' complex natural products, in large part "because the process of trial and error in this complex chemical space is very slow due to barriers to synthesis".

### HUMAN ENDEAVOUR

Chemistry, then, shares a great deal with conventional manufacturing: it changes through innovations in design and fabrication. We don't make cars or

televisions the way we used to, so why should molecules be any different? We need to avoid romanticizing an imagined bygone age, as the designer William Morris harkened back to the folk crafts of a fictitious Middle Ages.

Better than making molecules more complicated or larger is making them more useful, and making them in more useful ways. Like architecture, chemistry deals in elegance in both design and execution. There has not been enough discussion of these aspects of the science: how they are manifested, how they motivate, how much they are worth conserving.

In contemplating automated synthesis, for example, a comparison from mathematics comes to mind. There is debate over whether a mathematical proof should be celebrated for its own sake, regardless of method, or for its elegance and form — how it was done. Does 'proof by machine' count? Such questions go to the heart of science as a human endeavour. We tell ourselves that the goals are knowledge and capability. But there are other things we value in it too. ■

**Philip Ball** is a freelance writer. His latest book is *Invisible: The Dangerous Allure of the Unseen*.

e-mail: [p.ball@btinternet.com](mailto:p.ball@btinternet.com)

1. Fukaya, K. *et al. Org. Lett.* **17**, 2570–2573, 2574–2577 (2015).
2. Zou, Y. *et al. J. Am. Chem. Soc.* **137**, 7095–7098 (2015).
3. Parr, B. T., Economou, C. & Herzon, S. B. *Nature* **525**, 507–510 (2015).
4. O'Connor, M. (ed.) *Pointers & Pathways in Research 41* (CIBA of India, 1963).
5. Nicolaou, K. C., Vourloumis, D., Winssinger, N. & Baran, P. S. *Angew. Chem. Int. Edn* **39**, 44–122 (2000).
6. Whitesides, G. M. *Angew. Chem. Int. Edn* **54**, 3196–3209 (2015).
7. Kleiner, R. E., Dumelin, C. E. & Liu, D. R. *Chem. Soc. Rev.* **40**, 5707–5717 (2011).
8. Kanan, M. W., Rozenman, M. M., Sakurai, K., Snyder, T. M. & Liu, D. R. *Nature* **431**, 545–549 (2004).
9. Li, J. *et al. Science* **347**, 1221–1226 (2015).
10. Gillis, E. P. & Burke, M. D. *J. Am. Chem. Soc.* **129**, 6716–6717 (2008).

For a list of further reading on this topic, see [go.nature.com/xrsdms](http://go.nature.com/xrsdms).





In Medellín, Colombia, water tanks are being repurposed to create public spaces that offer classrooms, cafes and theatres.

#### URBAN STUDIES

# Blueprint for a cooperative city

Colin Ellard examines a study of the new urban paradigm that fosters ‘deep sharing’.

As urbanizing countries grapple with the need to provide sustainable energy and transport for their burgeoning cities, start-up companies are creating a culture and economy of sharing. Many are commercial. The global home-rental ‘community’ Airbnb, for instance, has an estimated 60 million users in 34,000 cities. US-based transport company Uber, which links registered drivers with passengers by way of smartphones, is active in more than 360 cities across 6 continents. Car-sharing services such as Zipcar are also widespread, attracting millennials who blanch at the costs of car ownership (environmental as well as financial).

Commercially mediated sharing can have a dark side. Sharing can skew local economies. Property owners turning to Airbnb may convert entire buildings to de facto hotels in cities such as New York, potentially contributing to housing crises. And Uber uses a surge-pricing algorithm to match supply and demand, meaning that users can face unpredictably high fares during periods of peak demand.

There is an alternative: bottom-up ventures that are digital or based in communities, rather than commercial. In *Sharing Cities*, environmental consultant Duncan McLaren and urban-policy scholar Julian

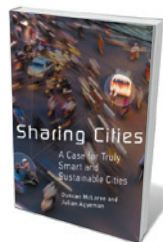
Agyeman lay out, with impressive depth, clarity and wisdom, a comprehensive prescription for a sharing paradigm that incorporates such models. Noting that sharing has been a sociocultural and informal practice for millennia, McLaren and Agyeman also reveal the promise and pitfalls of such an approach at a time when neoliberal economic policies emphasizing individualistic profit often trump public goods and services.

*Sharing Cities* explores the potential in dense urban spaces for ‘deep sharing’ of goods, resources, services, talent and experience through the Internet, with its rapid, extensive linking of lenders and borrowers. The models that the authors examine include barter clubs, credit unions, cooperative land trusts and co-housing to online and other peer-to-peer (P2P) networks and supper clubs. As they point out, commercial services barely scratch the surface of what is possible in a true

sharing economy. For example, decentralized P2P networks such as TaskRabbit — in which users can exchange skills and services without strong corporate oversight — can facilitate substantial sharing networks with minimal supervision. Public-transport systems can be considered a form of sharing, because the costs of mobility are shared between many.

Each chapter focuses on a particular aspect of sharing (production, consumption, politics, justice), and opens with a vignette of a city that exemplifies it. San Francisco, California — a hotbed of entrepreneurial start-up culture driven in part by Silicon Valley — is used to illustrate consumption. That kick-starts a discussion of open skills and knowledge sharing on online collaborative platforms that fly in the face of conventional commercial secrecy.

Medellín, Colombia, is used as an example of sharing in the context of social justice, as a result of the city’s spectacularly successful overturning of social marginalization over the past decade. This has been achieved through an ongoing architectural transformation of water tanks into shared public spaces, as well as the introduction of its



**Sharing Cities:**  
A Case for Truly  
Smart and  
Sustainable Cities  
DUNCAN MCLAREN  
AND JULIAN AGYEMAN  
MIT Press: 2016.



sustainable Metroplus bus rapid transit system. McLaren and Agyeman describe how other cities can foster inclusivity and sharing through prudent adjustments in policy and priorities, provision of open data and more thoroughgoing input from citizens at the grass-roots level.

As I worked my way through each chapter, I rode a crest of optimism about the imminence of real change, only to crash back to reality as I realized how difficult it is to ensure that sharing transformations are transparent, equitable and just. The authors never flinch from tackling the complexities and contradictions inherent in these examples. They present exquisitely balanced explanations of both the potential of sharing and its vulnerability to corruption by opportunistic invaders seeking to maximize profit over fairness.

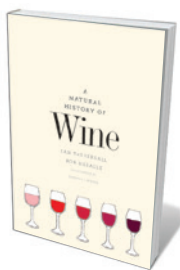
In many cases, as McLaren and Agyeman show, overcoming conflicts between bottom-up and profit-driven sharing ventures demands reconfiguration of urban policy. Two examples of this are participatory budgeting, in which citizens share responsibility for allocating resources, and shared land ownership, which emphasizes a public commons. Both deter the exclusions often generated by gentrification.

My only criticism is with one of the book's key premises: that humans are evolutionarily predisposed to share across the board. The authors point to work in developmental psychology showing that babies are aware of fairness and injustice (M. F. H. Schmidt and J. A. Sommerville *PLoS ONE* 6, e23223; 2011). Yet there is no shortage of evidence in evolutionary psychology — and everyday life — for the human tendency towards selfishness under some circumstances, towards some classes of others. And theoretical work has suggested that under many conditions common in human society, cooperation is likely to collapse (A. J. Stewart and J. B. Plotkin *Proc. Natl Acad. Sci. USA* 111, 17558–17563; 2014). Indeed, even the cited work by Schmidt and Sommerville shows that more than one-third of the infants in the study kept the best 'loot' for themselves.

In part, such differences are surely what underlie the constant push–pull between new sharing paradigms and the ventures that co-opt and parasitize them. It would have helped the balance of McLaren and Agyeman's argument to describe some of the seamy underbelly of our evolutionary heritage as well as the rosier side of our natures. ■

**Colin Ellard** is a cognitive neuroscientist at the University of Waterloo in Canada, specializing in the study of the relationship between human psychology and urban design. His latest book is *Places of the Heart*. e-mail: cellard@uwaterloo.ca

## Books in brief



### A Natural History of Wine

Ian Tattersall and Rob DeSalle YALE UNIVERSITY PRESS (2015)

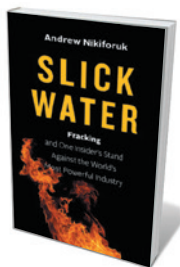
Was science ever more intoxicating? This sparkling contribution to the science of wine by palaeoanthropologist Ian Tattersall and entomologist Rob DeSalle draws on a staggering array of disciplines, from neurobiology to physics. Starting at the putative cradle of wine-making — an Armenian cave containing a 6,000-year-old proto-winery — the two trawl the research on frugivorous higher primates' putative hankering for fermented fruit; the bodily journey of a "wine-derived ethanol molecule"; and the impact of climate change on cultivation (J. Goode *Nature* 492, 351–353; 2012).



### White Eskimo: Knud Rasmussen's Fearless Journey into the Heart of the Arctic

Stephen R. Bown DA CAPO (2015)

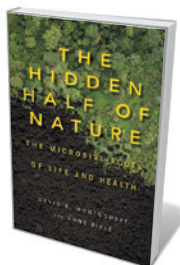
The part-Inuit, part-Danish explorer Knud Rasmussen is famed for his 32,000-kilometre Fifth Thule Expedition (1921–24) from Hudson Bay to Alaska. But as Stephen Bown reveals in this masterful biography, he was also an Arctic Richard Francis Burton, publishing key anthropological works on Inuit culture in Canada and Greenland. Ultimately, Bown shows, Rasmussen became a scientist-bohemian "as comfortable in bearskin pants on a featureless wind-lashed plain as he was in a formal suit and bow tie attending the opera".



### Slick Water: Fracking and One Insider's Stand against the World's Most Powerful Industry

Andrew Nikiforuk GREYSTONE (2015)

This meticulously researched study by journalist Andrew Nikiforuk lifts the lid on the costs of that vast geological-engineering experiment, fracking. It centres on Canadian environmental impact assessor Jessica Ernst, who in 2005 found explosive levels of methane in her well water, fingered the culprit as fracking and launched a legal battle. Interwoven with her story is a deft history of fracking from the 1850s (when torpedoes and nitroglycerin were used) through the 1960s (nuclear explosions) to modern hydraulic fracturing.



### The Hidden Half of Nature: The Microbial Roots of Life and Health

David R. Montgomery and Anne Biklé W. W. NORTON (2015)

Soils and the human gut teem with microbes, and both communities need care and feeding to support, respectively, nutrient-rich crops and healthy immune systems. So emphasize geologist David Montgomery and biologist Anne Biklé in this beautifully synthesized scientific memoir. Personal experiences — revitalizing degraded garden soil and surviving a major health scare — become ways into swathes of cutting-edge research in microbiology, from agronomist Lorenz Hiltner's work on "disease suppressive" soils to the Human Microbiome Project (see [go.nature.com/tsty3t](http://go.nature.com/tsty3t)).



### The Snowflake: Winter's Frozen Artistry

Kenneth Libbrecht and Rachel Wing VOYAGEUR (2015)

In 2003, physicist Kenneth Libbrecht (*J. Hoffman Nature* 480, 453–454; 2011) published the first edition of this aesthetic and scientific celebration of the snowflake. With park ranger Rachel Wing, Libbrecht returns with fresh research, more advanced microphotographs and a history of snowflake imaging from Robert Hooke's 1665 drawings to Wilson Bentley's photographs, taken between 1885 and 1931. A gallery of jewels — Antarctic 'diamond dust', roccoco stellar dendrites and beyond. **Barbara Kiser**

# In the cross hairs of controversy

Nancy Baron reviews a handbook for scientists keen to influence policy.

In the world of science communication, academia has its own scarlet letter: A, for Advocacy. Many scientists shudder at the thought of being branded advocates. As a result, they can undermine the message of their research by caveating every assertion, or even avoiding interaction with the public — something I have encountered many times as a science-communication coach.

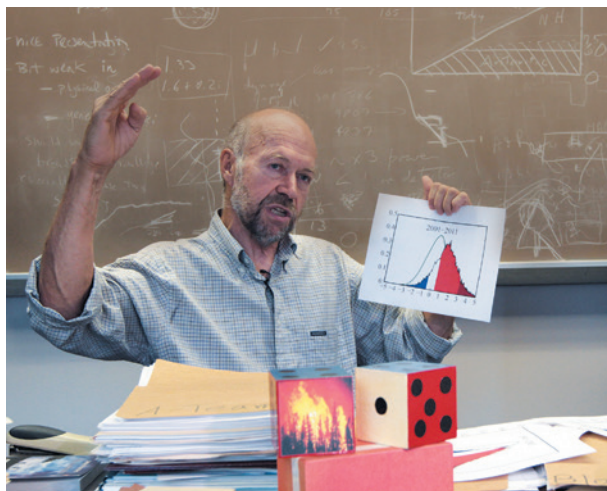
Lee Badgett is an academic who has fought with that tension and come out swinging. A professor of economics and policy, she is a veteran analyst and public intellectual with decades of experience in policy debates about equality for lesbian, gay, bisexual and transgender (LGBT) people. Her pithy Twitter profile describes her life's work as "Studying LGBT economic inequality to figure out how to end it." *The Public Professor*, her third book, is an exhortation to scientists to become "activist-scholars" like her.

Badgett intends to reverse-engineer advocate-academics to teach others how to galvanize policymakers, the media and the community to pay attention to their research. Her prescription for confronting injustice in areas from civil rights to climate change entails "injecting scholarship into important debates, taking advantage of good timing, being willing to handle disagreement" and connecting with the public, activists and policymakers.

Badgett breezes past reasons not to engage, averring that as for bad news about advocacy, "There really isn't any." As someone who entered the fray as a union organizer during her graduate degree, she draws on her own experience to offer strategies for researchers to inform legislative chambers, courtrooms, businesses and social movements.

She recommends three steps for maximum impact. First, examine the big picture, understand the debate and master the rules of the game, including determining your own role in a conversation. If an economist wants to recommend changes to the minimum wage, for example, understanding the institutions that regulate wage policy is essential. Identifying the decision-makers, as well as what they need and when, is crucial in working out how to make your research relevant and timely.

Second, build a network in the social spheres you hope to influence. Badgett advises using your existing network for e-mail introductions, or finding a hook to make a



Climate scientist and activist James Hansen.

cold call. A lawmaker who has introduced legislation may be keen to hear about how your research is relevant to it.

Third, practise the art of communicating with people outside your sphere. Prepare an elevator speech: what would you say to a US Congress member if you were in a lift with them for 30 seconds? Legislators are notorious for their short attention spans. The key is to distil your insights in a way that highlights their importance to the targeted person.

The payback for all such efforts, Badgett rightly notes, is that they generate feedback and ideas that can inform future research questions and improve your teaching.

Badgett advises academics who work on hot-button issues such as gay marriage, minimum wage or climate change to learn to manage conflict by seeing what lies behind it. Often it is politics, not science. The cost of avoidance, she writes, "is allowing others to dictate the debate and public outcome". Nor, she says, is being in the thick of the argument as difficult as neophytes imagine. In an intriguing section, "Developing a Thicker Skin", she writes: "for some scholars who haven't yet dipped their toe into the sea of engagement, once you are all the way in, you'll get used to the temperature". The best defence against attacks is to live and work by basic, ethical principles, she advises.

**The Public Professor: How to Use Your Research to Change the World**

M. V. LEE BADGETT  
New York University  
Press: 2015.

Her discussion of "sustainable engagement" in the long term may prove especially helpful to those with fears of a

career-crushing backlash from going public. Badgett suggests addressing those concerns head on, and sooner rather than later. She does not fear the potential pitfalls, including obstacles to getting tenure, that can be a concern outside the social sciences. In ecology, for instance, public engagement can be viewed as a distraction from publishing — perhaps a different reaction from that in sociology or economics.

Badgett's tenure portfolio included a note from Barney Frank, long-time Massachusetts congressional representative, thanking her for sending him an article that he entered into the *Congressional Record*. She recommends building a network of academics who can vouch for your contributions to public debates and

convince a tenure committee that they are worthy of credit. I have seen this strategy work, but it is patchy and dependent on leadership within departments and institutions.

I enjoyed reading about Badgett's experience, and would have welcomed more on lessons she has learned in the cross hairs. Her book skates over the surface of a large pond and sometimes feels on thin ice, with too few in-depth examples. Nor does it reference what I consider core reading, such as Cornelia Dean's *Am I Making Myself Clear?* (Harvard Univ. Press, 2012), an elegant book on the fundamentals of talking to journalists; Dennis Meredith's *Explaining Research* (Oxford Univ. Press, 2010), which is like having a top-notch public-information officer assigned to you; and Randy Olson's *Houston, We Have a Narrative* (Univ. Chicago Press, 2015; see *Nature* 526, 321; 2015), an astute take on how to make science resonate through storytelling.

*The Public Professor* pushes the boundaries for scientists thinking of taking the public plunge. It will also be instructive to the more restrained scientist, as defined by Roger Pielke in *The Honest Broker* (Cambridge Univ. Press, 2007; A. A. Rosenburg *Nature* 448, 867; 2007). Researchers may feel that their fields are more constrained than the social-justice issues that Badgett champions, but *The Public Professor* has much to offer by exploring what is possible for those who want to change the world. ■

Nancy Baron is director of science outreach for communication-training service COMPASS and author of *Escape from the Ivory Tower*. e-mail: [nbaron@compassonline.org](mailto:nbaron@compassonline.org)

MARY ALTATTA/AF/PA IMAGES



# Correspondence

## Archives and citation miss equal authors

It is now common practice to include 'equal contributions' footnotes in papers that have multiple first or senior authors. Unfortunately, this information is not preserved by the archiving and citation processes. The omission diminishes the roles of equal but subsequently listed authors, discouraging scientists from working in collaborations and teams — the backbone of modern scientific progress.

Equal-authorship details can currently be found only in the papers themselves. These details are not available on indexing sites or in referenced citations, which are increasingly the main source of information for literature searches.

To rectify this oversight, indexers need publishers to code author-status information in a standard format. For example, journals could include an asterisk beside authors' names to indicate equal contributions for article-citation purposes. We call on all journals and indexers such as PubMed, Google Scholar and Thomson Reuters Web of Science to update their systems to reflect shared authorship.  
**Brian D. Brown, Miriam Merad**  
*Icahn School of Medicine at Mount Sinai, New York, USA.*  
[brian.brown@mssm.edu](mailto:brian.brown@mssm.edu)

## Design buildings for rapid evacuation

In today's terrorism-prone world (see, for example, *Nature* **528**, 7–8; 2015 and *Nature* **528**, 20–21; 2015), it is becoming increasingly important to ensure that buildings are designed to be speedily evacuated in an emergency.

Evacuation modelling is a relatively new field that uses computational tools to predict human behaviour in a stricken building. Algorithms represent the range of people's possible reactions in the event

of such a disaster (see, for example, E. D. Kuligowski *et al.* US National Institute of Standards and Technology Technical Note 1680; 2010). Models provide information on optimal evacuation strategies and allow buildings to be tested using real and hypothetical evacuation scenarios.

Making evacuation modelling mandatory in the design and assessment of existing and planned buildings that could be at risk would minimize the impact of attacks on occupants.

**Enrico Ronchi**  
*Lund University, Sweden.*  
[enrico.ronchi@brand.lth.se](mailto:enrico.ronchi@brand.lth.se)

## Common doctorates across Europe

The German medical doctorate system is not the only element that needs changing to overcome the ills that you discuss (see *Nature* **527**, 7; 2015). In our view, a European approach offers the best cure.

We suggest that Germany's medical degree should be modified to lead to a common European medical qualification: the vocational degree of Doctor of Medicine. Postgraduate medical research should be part of a different common European qualification: the academic degree of Doctor of Philosophy.

Scientific quality would be guaranteed if the criteria for attaining degrees were to be standardized across Europe, and if specialist postgraduate medical colleges were widely set up. Students and clinicians would then also be able to pursue their divergent scientific interests more easily.

A European core curriculum devised along these lines would reduce excessive pressure on students, enhance the mobility of students and graduates, and foster the growth of excellent health care and science. The International Federation of Medical Students' Associations and the European Medical

Students' Association have already laid the foundations for such a curriculum (see J. Hilgers *et al.* *Med. Teach.* **29**, 270–275; 2007).

**Stefan U. Hardt, Jannis Papazoglou, Benedikt W. Pelzer**  
*European Medical Students' Association, Brussels, Belgium.*  
[pmo@emsa-europe.eu](mailto:pmo@emsa-europe.eu)

## Clean energy enters virtuous cycle

Governments promised on 30 November to almost double global funding for clean-energy research (see [go.nature.com/n4qdsu](http://go.nature.com/n4qdsu)). Meanwhile, the very act of deploying emissions-cutting technologies to meet countries' climate pledges at the recent United Nations summit in Paris is likely to spur major innovation.

Such technological advances mean that cutting emissions can drive down the cost of further cuts in emissions (see [go.nature.com/j8ueaj](http://go.nature.com/j8ueaj)). For example, the price of photovoltaic modules for solar energy has fallen by 85% since 2000 as markets have grown; electricity costs from wind are now comparable to those from coal; and energy-storage technologies are improving.

Publicly funded research and development, early investment by the private sector, and efficient deployment are all crucial for innovation. Market growth in renewable energy is largely driven by government policies, which have unleashed private companies' research ingenuity and achieved economies of scale and greater productivity (see also J. E. Trancik *Nature* **507**, 300–302; 2014).

Recognizing the mutual reinforcement of cutting emissions and improving clean energy is essential for negotiating a long-term, ambitious climate deal. As global efforts add up, falling costs should allow for an international agreement to phase in emissions cuts at a rate that

matches each nation's stage of economic development.

**Jessika E. Trancik**  
*Massachusetts Institute of Technology, Cambridge, USA.*  
[trancik@mit.edu](mailto:trancik@mit.edu)

## Crowdfunded trials doubly scrutinized

We disagree that crowdfunding of clinical trials is ethically questionable (P. Y. Cheah *Nature* **527**, 446; 2015). Participants are still governed by the same high standards of research integrity as traditionally funded recipients — but with the added scrutiny that comes with public engagement (see, for example, N. Siva *Lancet* **384**, 1085–1086; 2014).

Cheah criticizes crowdfunding of clinical trials because it risks backing studies that are of limited importance and applicability. However, it is this very feature that offers an opportunity to fund trials for rare or emerging tropical diseases that might not otherwise attract financial support (see, for example, T. S. van der Werf *et al.* *Bull. World Health Organ.* **83**, 785–791; 2005).

**David Hawkes**  
*University of Melbourne; and Florey Institute of Neuroscience and Mental Health, Victoria, Australia.*

**Melanie Thomson**  
*Deakin University, Victoria, Australia.*  
[m.thomson@deakin.edu.au](mailto:m.thomson@deakin.edu.au)

### CORRECTION

The Outlook article 'Research without prejudice' (*Nature* **525**, S12–S13; 2015) incorrectly stated that the approval of the US National Institute on Drug Abuse (NIDA) is required for US cannabis trials. In fact, NIDA provides cannabis for every project that has completed the government-mandated approval process. The article also implied that NIDA was holding up the start of a trial led by Sue Sisley, but the delay is caused by other circumstances.



## BIODIVERSITY

## Recovery as nitrogen declines

Pollution from atmospheric nitrogen deposition is a major threat to biodiversity. The 160-year-old Park Grass experiment has uniquely documented this threat and demonstrated how nitrogen reductions lead to recovery. [SEE LETTER P.401](#)

DAVID TILMAN & FOREST ISBELL

Although greater availability of a scarce nutrient might seem beneficial for all plant species, this is not so. Even in seemingly pristine and protected ecosystems, large losses of plant diversity can be caused by the addition of a nutrient that limits plant growth<sup>1</sup>. One documented example is the effect of deposition on land of nitrogen that was released into the atmosphere by fossil-fuel combustion and agriculture. However, it has been unclear whether plant diversity will recover when nitrogen emissions are reduced or whether additional restoration practices are required. On page 401 of this issue, Storkey *et al.*<sup>2</sup> use the unparalleled long-term data of the Park Grass experiment to show that plant diversity recovers as nitrogen deposition decreases.

The Park Grass experiment at Rothamsted Research in Harpenden, UK, was started in 1856 and is the longest-running study of grassland in the world (Fig. 1). By comparing fertilized and control (never fertilized) plots, the authors observed that plant diversity declined to about 30% of its original level during 135 years of nitrogen fertilization, but returned to about 70% of its original level two decades after fertilization was halted. Moreover, plant diversity declined in unfertilized control plots to about 50% of its original level as atmospheric nitrogen deposition increased from 1950 to 1985. Then, when the introduction of cleaner technologies greatly decreased nitrogen deposition from 1985 to 2012, plant diversity increased to about 80% of its original level. In both recoveries, plant communities tended to regain their former species compositions.

These observations contrast with results of a grassland experiment in Minnesota in which little, if any, recovery had occurred two decades after cessation of high rates of nitrogen fertilization<sup>3</sup>. Storkey and collaborators suggest the intriguing possibility that this difference is due to the fact that the Park



**Figure 1 | The Park Grass experiment.** This field experiment in Hertfordshire, UK, has been running since 1856. Its division of plants into control or treated plots has been used to test the effects of various interventions on agricultural productivity, such as fertilization and altered soil pH. Storkey *et al.*<sup>2</sup> used data from the experiment to document declines in plant biodiversity in response to nitrogen accumulation, but also found that diversity recovers as nitrogen levels decrease.

Grass plots have been hayed (the grass cut, dried and removed) twice each year since the experiment started, whereas the Minnesota plots were never hayed. Why might this matter? Haying removes biomass and its nitrogen. If not removed, excess nitrogen that had accumulated in an ecosystem would recycle within that system, thereby retaining its ecological impacts long after nitrogen addition slowed or ceased.

Park Grass hay contains 1.5–2% nitrogen (see Fig. 2 of the paper<sup>2</sup>), and so annual removal of around 2 and 5 tonnes per hectare of hay from the control and fertilized plots, respectively,

probably removed around 35 and 90 kilograms of nitrogen per hectare per year. This removal occurred alongside a reduction in nitrogen deposition of around  $25 \text{ kg ha}^{-1} \text{ yr}^{-1}$  from its peak, for the control plots, and of an additional reduction of  $96 \text{ kg ha}^{-1} \text{ yr}^{-1}$  for the plots that stopped receiving fertilizer. The removals of nitrogen through haying possibly hastened the plots' recovery.

The reason why increased availability of limiting nutrients can cause biodiversity losses lies in the evolutionary trade-offs that cause species to be specialists. Adaptations that increase the ability of a given plant species to compete for one limiting resource come at a cost to the species' capacity to deal with limitation by another resource or factor<sup>4</sup>. Thus, an increased supply of one resource should lead to the competitive displacement of those species that are superior competitors for the enriched resource, because they are among the poorer competitors for the new limiting factor. In theory, if enrichment led to accumulation of a formerly limiting nutrient, both cessation of its addition and decreases in its stores would be needed for that nutrient to again become limiting and for biodiversity to begin recovering.

Although further tests are needed to confirm that haying helped the Park Grass plots' recovery, the idea is supported by several other examples. Human activities release more available nitrogen and phosphorus than all natural terrestrial processes combined<sup>5–8</sup>, and accumulation of these nutrients can cause dramatic shifts in species compositions and biodiversity in terrestrial and aquatic ecosystems. For instance, high-diversity heathlands in the Netherlands and Germany were replaced by low-diversity grasslands as nitrogen deposition reached higher rates than those in Britain<sup>9,10</sup>. Successful restoration of these heathlands often required physical removal of both vegetation and topsoil<sup>11–14</sup>. Similarly, 40 square kilometres of vegetable farmland in south Florida became a virtual monoculture of Brazilian pepper trees after it became part of Everglades National Park and agriculture

ROTHAMSTED RESEARCH

ceased in 1975. Attempts to restore the pre-agricultural ecosystem were futile until both the invasive trees and the fertilized agricultural soil were removed<sup>15–17</sup>. For phosphorus-limited lake ecosystems, reduction of phosphorus inputs can be insufficient for lake recovery if excess phosphorus inputs from agriculture are retained and recycled<sup>18</sup>.

These cases suggest that both reduction of nutrient inputs and removal of any large stores of accumulated nutrients may be required for restoration of native ecosystems. Some terrestrial restorations also require liming to overcome soil acidification, and seed addition when formerly abundant plant species are absent<sup>2,11,13</sup>. However, it is not yet clear how the magnitude of increases in nitrogen stores influences the recovery of grassland diversity after nitrogen addition decreases or ceases<sup>3,19,20</sup>.

The insights from the Park Grass experiment, together with results from earlier studies, show that biodiversity can recover even after chronic high rates of nutrient pollution, and suggest that this recovery may be hastened by, or perhaps require, management practices that reduce accumulated nutrient stores. Moreover, it suggests that haying, a much gentler practice than destructive removal of both vegetation and soil, may reduce nutrient stores sufficiently to allow grassland diversity to recover. Finally, Storkey and colleagues' work demonstrates the great value that long-term studies can provide in identifying solutions to environmental problems. ■

**David Tilman and Forest Isbell** are in the Department of Ecology, Evolution and Behavior, University of Minnesota, St Paul, Minnesota 55108, USA. **D.T.** is also in the Bren School of Environmental Science and Management, University of California, Santa Barbara.  
e-mails: [tilman@umn.edu](mailto:tilman@umn.edu); [isbell@umn.edu](mailto:isbell@umn.edu)

1. Harpole, W. S. & Tilman, D. *Nature* **446**, 791–793 (2007).
2. Storkey, J. *et al.* *Nature* **528**, 401–404 (2015).
3. Isbell, F., Tilman, D., Polasky, S., Binder, S. & Hawthorne, P. *Ecol. Lett.* **16**, 454–460 (2013).
4. Tilman, D. *Am. Nat.* **178**, 355–371 (2011).
5. Vitousek, P. M. *et al.* *Ecol. Appl.* **7**, 737–750 (1997).
6. Vitousek, P. M., Mooney, H. A., Lubchenco, J. & Melillo, J. M. *Science* **277**, 494–499 (1997).
7. Smith, V. H., Tilman, D. & Nekola, J. C. *Environ. Pollut.* **100**, 179–196 (1999).
8. Smil, V. *Annu. Rev. Energy Environ.* **25**, 53–88 (2000).
9. Bakker, J. P. *Nature Management by Grazing and Cutting* (Kluwer, 1989).
10. Bakker, J. P. & Berendse, F. *Trends Ecol. Evol.* **14**, 63–68 (1999).
11. Aerts, R., Huiszoon, A., Van Oostrum, J. H. A., Van De Vijver, C. A. D. M. & Willems, J. H. J. *Appl. Ecol.* **32**, 827–835 (1995).
12. Jansen, A. J. M., de Graaf, M. C. C. & Roelofs, J. G. M. *Vegetatio* **126**, 73–88 (1996).
13. Dorland, E. *et al.* *Plant Soil* **265**, 267–277 (2004).
14. Niemeyer, M., Niemeyer, T., Fottner, S., Härdtle, W. & Mohamed, A. *Biol. Conserv.* **134**, 344–353 (2007).
15. Li, Y. & Norland, M. *Soil Sci.* **166**, 400–405 (2001).
16. Dalrymple, G. H., Doren, R. F., O'Hare, N. K.,

- Norland, M. R. & Armentano, T. V. *Wetlands* **23**, 1015–1029 (2003).
17. Smith, C. S., Serra, L., Li, Y., Inglett, P. & Inglett, K. *Crit. Rev. Environ. Sci. Technol.* **41**, 723–739 (2011).
18. Carpenter, S. R., Ludwig, D. & Brock, W. A. *Ecol. Appl.* **9**, 751–771 (1999).

19. Fornara, D. A. & Tilman, D. *Ecology* **93**, 2030–2036 (2012).
20. Clark, C. M. & Tilman, D. *Ecology* **91**, 3620–3630 (2010).

This article was published online on 2 December 2015.

## QUANTUM PHYSICS

# Entanglement beyond identical ions

Control of quantum particles has been extended to enable different types of ion to be entangled — correlated in a non-classical way. This opens up opportunities for the development of new quantum technologies. [SEE LETTERS P.380 & P.384](#)

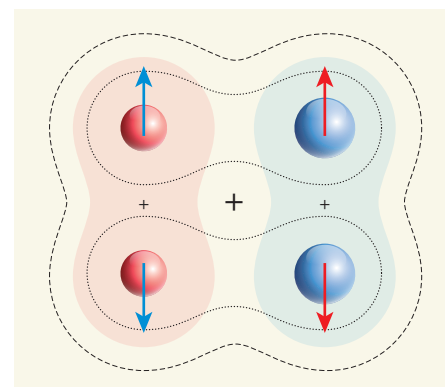
TOBIAS SCHAEZT

Entanglement is a peculiar phenomenon that causes two or more particles to share one common state, such that each particle can no longer be described independently. In this issue, Tan *et al.*<sup>1</sup> (page 380) and Ballance *et al.*<sup>2</sup> (page 384) report entangled pairs of ions consisting of two different atomic species — the first time that this has been achieved. They used the resulting systems to test the puzzling predictions of quantum mechanics with unprecedented accuracy. This in turn allowed them to benchmark trapped ions as an experimental platform for quantum technology, and to assess the platform's prospects to further exploit quantum effects for applications such as atomic clocks and quantum computation.

Quantum mechanics requires objects to be able to exist in two states simultaneously, even if the states are mutually exclusive. To picture such a superposition, imagine the magnetic needle of a hypothetical quantum compass pointing north and south at the same time. A measurement that determines the state of the needle will project it into one of its two possibilities at random — the result is not just unknown, but not determined before the measurement.

If there are two quantum magnetic needles, they can become entangled. For entangled objects, a measurement on one object that produces a completely random output instantaneously determines the potential result of the second object (or vice versa). The effect of the measurement is immediate and is independent of the distance between the objects.

Einstein was one of the founding fathers of the theory of quantum mechanics, but he and his colleagues realized that the consequences of entanglement severely violate intuition and logical conclusions based on the classical interpretation of nature. Einstein and others therefore proposed some seminal experiments<sup>3</sup> that could be used to show that their theory was far from complete. But because the practical



**Figure 1 | Entangling two different ions.**

Individual ions can exist in one of two quantum 'spin' states: spin up (↑) and spin down (↓). Quantum mechanics also allows ions to form a superposition state (↑↓) in which both the ↑ and ↓ states coexist. Tan *et al.*<sup>1</sup> and Ballance *et al.*<sup>2</sup> have prepared entangled pairs of ions consisting of two different atomic types — either different elements or different isotopes of an element. Each ion seems to be in a ↑↓ state (shaded regions), but entanglement generates a correlated state (↑↑↓↓), bounded by dashed lines; dotted lines indicate spin correlations), which means that a measurement of one of the two ions instantaneously affects the state of the other — that is, the two formerly independent ions have to be considered as a whole.

prerequisites for the experiments seemed to exceed the capabilities of any researcher, even in the future, they called their proposals *Gedankenexperimente* ('thought experiments').

Tan *et al.* and Ballance *et al.* report that quantum mechanics is accurate even when non-identical objects are entangled. Tan and colleagues entangled a beryllium-9 ion (<sup>9</sup>Be<sup>+</sup>) and a magnesium-25 ion (<sup>25</sup>Mg<sup>+</sup>), whereas Ballance and co-workers used two isotopes of calcium, <sup>40</sup>Ca<sup>+</sup> and <sup>43</sup>Ca<sup>+</sup>. To describe how both groups created entanglement, consider the ions in each pair as magnetic needles that can point in one of two directions. This behaviour is analogous to that of a particle that has



a spin value of  $+\frac{1}{2}$  or  $-\frac{1}{2}$ ; discrete spin values are a quantum form of angular momentum. Applying appropriate optical fields generated by laser beams, or microwave fields, mediates a ferromagnetic interaction that aligns the spins. In other words, if the first ion is prepared and kept in the 'northward-pointing' spin-up ( $\uparrow$ ) state, then the interaction puts the second spin into a  $\uparrow$  state too.

In a similar way, the authors prepared the first spin in a superposition state ( $\uparrow+\downarrow$ ) by switching off the spin-rotating microwave or laser fields after  $90^\circ$  of rotation. The researchers then induced the ferromagnetic interaction described above. This orients the second spin into an entangled superposition state of ferromagnetic order ( $\uparrow\uparrow+\downarrow\downarrow$ ); the  $\uparrow$  part of the first ion's superposition state rotates the second ion into  $\uparrow$ , and the  $\downarrow$  part rotates the second ion into  $\downarrow$  (Fig. 1). The quantum nature of the created correlation became evident when the researchers took measurements of only the first ion's spin. The outcome was completely random but instantaneously determined the outcome of a subsequent measurement of the second spin — the outcome of the second measurement was almost always identical to that of the first.

Some correlation of measurements of classical objects is possible, and this is potentially enhanced in the presence of unknown or hidden (but classical) variables. The maximal possible correlation by classical means can be derived mathematically in the form of an inequality, known as a Bell inequality. In the current experiments, the variety concerned is called the CHSH Bell inequality, and its upper bound for classically achievable correlations is 2. Entanglement requires quantum correlations that enable this upper bound to be exceeded — that is, the Bell inequality can be violated up to a maximum value of approximately 2.828. When such violations are measured experimentally, the results show that entanglement is necessary to describe nature.

In 1982, the first experimental tests were done<sup>4</sup>, and demonstrated that entanglement does indeed seem to be necessary. Since then, any potential shortcomings in the experiments used to find violations of Bell inequalities have been ruled out<sup>5,6</sup>, albeit within statistical error limits. Tan and colleagues report a violation of up to 2.70, with a residual uncertainty that essentially rules out any classical description of nature — their result is equivalent to being about 40 standard deviations away from the value obtainable using classical explanations. When preparation and readout errors in Ballance and co-workers' study are accounted for, the theoretical maximum of the Bell inequality is 2.236; the authors report a violation of 2.228, with an uncertainty that means that the value differs by 15 standard deviations from any classical description.

The results emphasize that science and engineering at the level of individual quanta can reveal and characterize quantum mechanics

with unprecedented accuracy, at close to 100% detection efficiency. But they also impressively demonstrate how the total quantum performance of a system can be benchmarked — the proximity of the experimentally determined violations to their theoretical limits quantifies the quality, performance or fidelity of the quantum operations in a single number.

The findings substantially improve the prospects for designing and realizing devices that use superposition states and entanglement as reliable resources, based on trapped ions or related systems. Different tasks in a common experimental protocol can now be allocated to the atomic species best suited for the chosen purpose — such as quantum memory, performance of logic operations with negligible effects on any nearby quantum memory elements, and generating links to devices based on other technological platforms, such as photonic or solid-state devices. This paves the

way for precise spectroscopy, ultra-accurate clocks and simulators of quantum systems. It might even enable the development of universal quantum computers capable of running a superposition of many correlated tasks in parallel, offering much better performance than is currently available using conventional computers, such as exponentially higher speeds for dedicated applications. ■

**Tobias Schaetz** is at the Institute of Physics, Albert Ludwig University of Freiburg, Freiburg 79104, Germany.  
e-mail: tobias.schaetz@physik.uni-freiburg.de

1. Tan, T. R. *et al. Nature* **528**, 380–383 (2015).
2. Ballance, C. J. *et al. Nature* **528**, 384–386 (2015).
3. Einstein, A., Podolsky, B. & Rosen, N. *Phys. Rev. A* **47**, 777–780 (1935).
4. Aspect, A., Grangier, P. & Roger, G. *Phys. Rev. Lett.* **49**, 91–94 (1982).
5. Rowe, M. A. *et al. Nature* **409**, 791–794 (2001).
6. Hensen, B. *et al. Nature* **526**, 682–686 (2015).

## REPRODUCIBILITY

# Experimental mismatch in neural circuits

**The finding that acute and chronic manipulations of the same neural circuit can produce different behavioural outcomes poses new questions about how best to analyse these circuits. SEE ARTICLE P.358**

THOMAS C. SÜDHOF

In 1949, Walter Rudolf Hess shared the Nobel Prize in Physiology or Medicine for his work using acute electrical stimulation to study neural circuits. Modern neuroscience is dominated by a newer, more sophisticated technique for acute circuit manipulation: optogenetics, in which light-sensitive ion-channel proteins are engineered to activate or inhibit select neurons<sup>1</sup>. However, a nagging doubt pervades the field — do the behavioural effects of acutely activating or silencing specific neurons reflect the normal functions of these cells? On page 358 of this issue, Otchy *et al.*<sup>2</sup> systematically address this question. Their findings are bound to excite lively discussion.

If acute inactivation of a particular neural circuit alters an animal's behaviour, the seemingly logical conclusion is that the circuit controls the behaviour. But the brain's circuits are densely interconnected, so how can we be sure that these behavioural effects are not caused by changes to other, connected, circuits that normally do not participate in the targeted behaviour but are affected by the manipulation? Otchy *et al.* used a brilliant study design to test this idea. They reasoned that, if the effects of acute manipulation are directly

caused by the manipulated neurons, then chronically manipulating those neurons, for example by permanently impairing (lesioning) them, should have the same effect. The authors compared the effects of chronic and acute neural manipulations in rats and in zebra finches. They examined behavioural tasks that were learnt before the manipulations, but that were not repeatedly practised afterwards, avoiding the confounding effect of relearning a task after an experimental manipulation.

First, Otchy *et al.* demonstrated that, in rats that had learnt a complex lever-pressing task, acute silencing of neurons in the brain's motor cortex using the drug muscimol profoundly impaired task performance. Acute optogenetic activation of motor-cortex neurons produced a similar effect. The same research group had shown previously<sup>3</sup> that surgical ablation of the motor cortex blocked the initial learning of the lever-pressing task, but had no significant effect on the ability of rats to perform the task if it had been learnt before surgery. Thus, acute and chronic manipulations produce discrepant results in this circuit (Fig. 1a).

In a second set of experiments, Otchy and colleagues used muscimol to inactivate song neurons in a brain region called the sensorimotor nucleus interface (Nif) in zebra

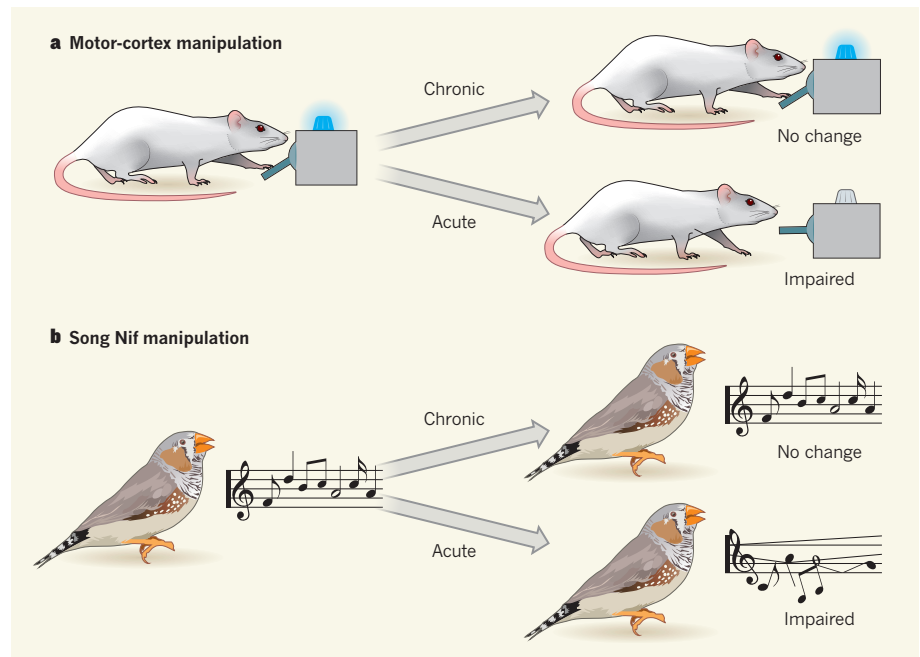
finches. This acute manipulation massively impaired birdsong, whereas chronic lesioning of Nif had no effect two days after the lesion (Fig. 1b). Investigating this apparent paradox, the authors showed that the Nif lesions did initially cause a change in the downstream neural circuitry controlling birdsong, but that this change spontaneously recovered without training after 3.4 hours. The researchers propose that homeostatic plasticity, which adjusts the overall activity level of neurons in a circuit, might be involved in this recovery. However, other processes that change the strength of the synaptic connections between these neurons are equally likely to be responsible.

How should we interpret these experiments? Two opposing hypotheses come to mind. First, that acute manipulations are unreliable and should be discarded in favour of chronic manipulations. Second, that acute manipulations elicit results that truly reflect normal circuit functions, and the lack of changes after chronic manipulations is caused by compensatory plasticity.

Before choosing between these stark alternatives, several facts should be taken into account. Many chronic manipulations of neural circuits (both permanent genetic changes and physical lesions) do actually produce major behavioural changes. For example, in rodent and human brains, lesions in the amygdala region impair fear memories<sup>4</sup>, and hippocampal lesions interfere with spatial memory<sup>5</sup>. Chronic deletion of the synaptic cell-adhesion molecule neuroligin-3 in striatal neurons alters learning of a repetitive motor task<sup>6</sup>. Thus, the finding that a chronic manipulation does not cause a behavioural change cannot simply be attributed to plasticity and compensation.

Clearly, it is possible to dissect the functions of some types of neuron and circuit using chronic manipulations, making this a compelling overall experimental approach. But acute optogenetic manipulations are generally easier to perform, and the conclusions drawn from many such manipulations do correlate well with those from chronic manipulations (see, for example, ref. 4). Moreover, such acute manipulations often match changes in neural activity observed during the targeted behaviour *in vivo*<sup>7–9</sup>, although a caveat of acute manipulations is that natural neural activity is normally limited to only a subset of neurons in a circuit, whereas acute manipulations are mostly not.

There are multiple explanations for why acute and chronic manipulations might produce distinct results, which makes it difficult, or perhaps even impossible, to assess whether results reflect 'off-target' or 'on-target' effects, as Otchy *et al.* aptly call them. The authors point out that, because neural circuits are massively interconnected, acute manipulations are probably more susceptible to off-target effects than are chronic lesions. This is because acute manipulations are



**Figure 1 | Mixed messages from neural manipulations.** Otchy *et al.*<sup>2</sup> compared the effects of acute and chronic manipulations of neural circuits on a specific behaviour. **a**, The authors taught rats to perform a complex lever-pressing task. Chronic inhibition of neurons in the brain's motor cortex did not affect task performance, whereas acute perturbations strongly impaired performance. **b**, Likewise, chronic ablations of neurons in the sensorimotor nucleus interface (Nif) of the brains of zebra finches did not affect their songs, whereas songs became variable and unstructured after acute inhibition.

more likely to spread to other connected circuits that have no normal role in the targeted behaviour. Therefore, we cannot simply assume that the behavioural readouts of such manipulations always reflect the normal functions of the manipulated circuits.

Where do we go from here? Most acute manipulation studies that use optogenetics confirm, and so add valuable support to, existing hypotheses that were established in earlier studies. But for those studies that have proposed new circuit functions, it may be advisable to re-evaluate the conclusions using independent approaches.

In the future, it might be helpful always to correlate acute and chronic manipulations of specific neurons. If results from acute and chronic manipulations are discrepant, analyses of circuits that act in parallel to the manipulated circuit, or of similar neurons that are activated by different stimuli, might be more likely to provide an explanation for the discrepancy than examination of chains of hierarchically connected neurons, because off-target effects probably propagate throughout neural circuits by spilling over into adjacent, connected circuits. Moreover, studies of a broad range of behaviours might be helpful — restricting a study to a few behaviours could make it harder to detect off-target effects. Overall, more caution about the conclusions drawn from circuit manipulations, be they acute or chronic, seems advisable, because most current studies focus on only one circuit and one behaviour.

It is both an exciting and a sobering time for neuroscience. Exciting, because it is now possible to manipulate neurons and circuits with an ease that was only dreamt of a few years ago. Sobering, because the massively parallel and interconnected nature of neural circuits is becoming apparent, and the complexity imposed on such circuits by various forms of plasticity has yet to be even touched on. By using parallel approaches to study circuits, we can develop an understanding of the brain that acknowledges the limitations of this understanding, as well as its achievements. Such a strategy will drive the field forward. ■

**Thomas C. Südhof** is in the Department of Molecular and Cellular Physiology, and at the Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA.  
e-mail: tcs1@stanford.edu

- Emiliani, V., Cohen, A. E., Deisseroth, K. & Häusser, M. *J. Neurosci.* **35**, 13917–13926 (2015).
- Otchy, T. M. *et al.* *Nature* **528**, 358–363 (2015).
- Kawai, R. *et al.* *Neuron* **86**, 800–812 (2015).
- Johansen, J. P., Wolff, S. B., Lüthi, A. & LeDoux, J. E. *Biol. Psychiatry* **71**, 1053–1060 (2012).
- Squire, L. R., Genzel, L., Wixted, J. T. & Morris, R. G. *Cold Spring Harb. Perspect. Biol.* **7**, a021766 (2015).
- Rothwell, P. E. *et al.* *Cell* **158**, 198–212 (2014).
- Buetfering, C., Allen, K. & Monyer, H. *Nature Neurosci.* **17**, 710–718 (2014).
- Chaumont, J. *et al.* *Proc. Natl Acad. Sci. USA* **110**, 16223–16228 (2013).
- Kravitz, A. V., Owen, S. F. & Kreitzer, A. C. *Brain Res.* **1511**, 21–32 (2013).

This article was published online on 9 December 2015.



## MICROBIOMES

# Curating communities from plants

Large-scale cultivation and genome sequencing of the bacteria that inhabit the leaves and roots of *Arabidopsis* plants have paved the way for probing how microbial communities assemble and function. [SEE ARTICLE P.364](#)

GWYN A. BEATTIE

Vast networks of microorganisms live in our soils, seas and bodies. These microbiomes also develop in intimate association with plants, in which they can enhance nutrient uptake, growth and tolerance to pathogens, pests and environmental stresses. Recognition of the fundamental role of microbes in the health of plants and animals, and the centrality of microbes in many ecological processes, has led to recent proposals for international<sup>1</sup> and US-based<sup>2</sup> microbiome initiatives. These proposals have highlighted a key need to develop collections of cultured organisms for experimental enquiry into the function and assembly of native communities<sup>1</sup>. On page 364 of this issue, Bai *et al.*<sup>3</sup> describe genome-sequenced bacterial culture collections that represent most of the species in native root- and leaf-associated microbiomes of *Arabidopsis thaliana* plants. They show that these collections can be used to reproducibly establish communities that resemble those found naturally on wild plants.

High-throughput genomic sequencing is enabling the characterization of microbiome profiles based on nucleic-acid signatures and the total gene content of a community (metagenomics). The breadth and depth of this

profiling is increasing with the affordability of sequencing. Because this approach does not require the microorganisms to be cultivated, it has transformed our understanding of the taxonomic composition and gene content of animal- and plant-associated microbiomes. For example, cultivation-independent profiling of the root microbiota of the model flowering plant *Arabidopsis* has highlighted compositional consistencies not only across soils from multiple continents<sup>4,5</sup>, suggesting that these microbiomes share common assembly processes, but also across multiple *Arabidopsis* lineages<sup>6</sup>, suggesting their evolutionary conservation.

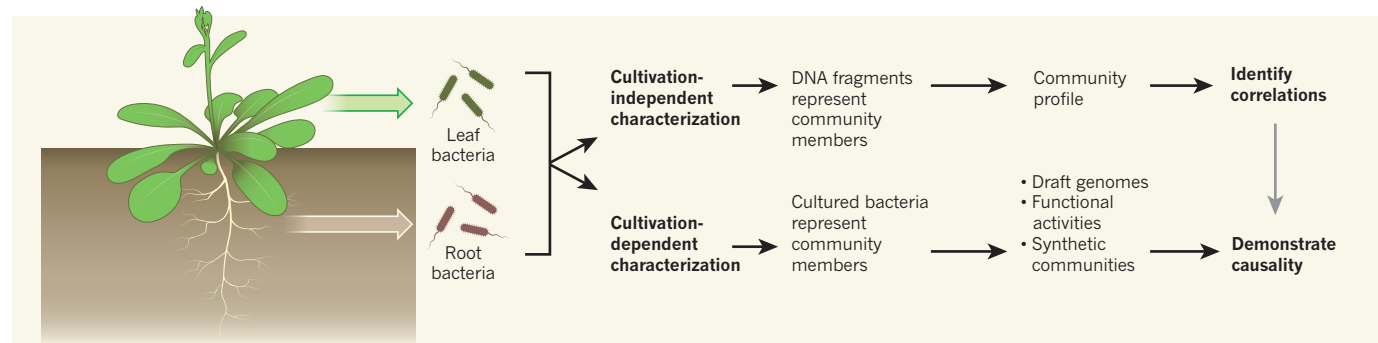
However, uncovering the microbiome assembly mechanisms requires the ability to manipulate microbial communities, including engineering and perturbing synthetic communities. Thus, experimental enquiry into microbiomes requires more than sequence data — it needs microbial cultures (Fig. 1).

Bai and colleagues amassed and identified almost 8,000 bacterial isolates from the roots and leaves of *Arabidopsis* plants grown in the field or in the laboratory in soils taken from the field. These collections included representatives of most bacterial species that have been identified in *Arabidopsis* microbiomes by cultivation-independent profiling<sup>4–8</sup>, which suggests that most bacteria

associated with *Arabidopsis* leaves and roots are readily cultivated. This culturability of plant-associated bacteria contrasts sharply with the historic inability to culture the vast majority of bacteria in soil and aquatic habitats<sup>9</sup>, and it probably results from root and leaf habitats being rich in organic compounds and oxygen. The finding that these communities can be so well represented by culture collections highlights the value of plant microbes as models for investigating the mechanisms of microbiome assembly and function.

Bacteria associated with the roots and leaves of terrestrial plants generally fall into only a few phyla that are shared between these plant tissues<sup>4–8</sup>. By generating taxonomically representative culture collections of microbes from roots (194 isolates) and leaves (206 isolates), Bai *et al.* established that bacterial families in these phyla are generally found on both tissue types. However, the function of microbiomes, particularly with regard to their impact on the host plant, is probably strongly rooted at the species, subspecies and strain level, and information at these levels is captured by sequencing whole genomes.

The authors generated high-quality draft genome sequences of their 400 root and leaf isolates, as well as 32 soil isolates, and examined how the phylogenetic and functional diversity among isolates within microbial families correlates with their origins in roots or leaves. They found some evidence for microbial specialization to either the leaf or root niche: a few phylogenetic clusters were found only or primarily in one niche, and certain functional characteristics — such as the degradation of foreign chemical substances — were enriched in one niche more than the other. However, the taxonomy of the isolates predicted their functional diversity much better than did their origins on roots or leaves. The authors' recognition of prominent family-level differences in functional



**Figure 1 | From correlation to causation.** Cultivation-independent profiling of microbial communities involves sequencing DNA fragments amplified from cells to generate a comprehensive picture of the community members. These profiles can be used to identify correlations, such as the presence of specific microbes on leaves versus roots, and to evaluate the extent to which culture collections represent the complete community. Bai *et al.*<sup>3</sup> generated large collections of bacteria associated with the leaves and roots of

*Arabidopsis thaliana* plants and found them to be highly representative of the species present in cultivation-independent profiles. The authors used these culture collections to derive draft genomes, evaluate potential functional activities and create synthetic communities that, when applied to initially microbe-free plants, allowed experimental evaluation of factors that drive the assembly of leaf- and root-associated microbial communities.

diversification demonstrates a need for studies into how distinct taxonomic groups contribute to microbiome function.

Synthetic microbial communities can be used to systematically query natural microbiome processes. Bai *et al.* introduced synthetic communities of 188 and 218 representative isolates from root (or soil) and leaf communities, respectively, onto gnotobiotic *Arabidopsis* plants — plants that were microorganism-free before inoculation with known microorganisms. They then evaluated the communities that assembled by sequencing genes that help to identify the taxa (the 16S ribosomal RNA genes). These synthetic communities yielded assemblages on gnotobiotic plants that had consistent compositions, showing reproducibility in microbiome assembly processes; moreover, their composition resembled the native bacterial microbiomes found on wild *Arabidopsis* plants. Surprisingly, the resulting communities were not influenced by the relative proportion of the applied strains, indicating that community assembly is a robust process.

The synthetic communities were also instrumental in teasing apart two of the drivers of community assembly on *Arabidopsis* leaves: the source of the isolates (roots or leaves), and their arrival through the air or the soil. These findings demonstrate how synthetic communities can serve as windows on the origins and development of the bacterial component of plant microbiomes.

We are at a crucial juncture in microbiome research, transitioning from cataloguing microbes and genes to executing hypothesis-driven experiments. Bai and colleagues have provided resources that will speed this transition for plant research, including a

large culture collection, complex synthetic communities with sequenced genomes and a gnotobiotic reconstitution system. Together, these resources enable recapitulation of the assembly of native bacterial communities on *Arabidopsis* plants, facilitating studies that provide ecologically relevant answers to questions about the establishment, dynamics, resilience, function and evolution of plant microbiomes. The mechanistic understanding derived from these synthetic communities is an excellent step on the road to understanding how the sustained health and productivity of our agricultural and natural systems are influenced by plant microbiomes and, more broadly, by phytobiomes — the networks of bacteria, fungi, oomycetes, viruses, nematodes, insects and other animals that affect plants. ■

Gwyn A. Beattie is in the Department of Plant Pathology and Microbiology, Iowa State University, Ames, Iowa 50014-3211, USA.  
e-mail: gbeattie@iastate.edu

1. Dubilier, N., McFall-Ngai, M. & Zhao, L. *Nature* **526**, 631–634 (2015).
2. Alivisatos, A. P. *et al. Science* **350**, 507–508 (2015).
3. Bai, Y. *et al. Nature* **528**, 364–369 (2015).
4. Bulgarelli, D. *et al. Nature* **488**, 91–95 (2012).
5. Lundberg, D. S. *et al. Nature* **488**, 86–90 (2012).
6. Schlaeppi, K., Dombrowski, N., Oter, R. G., van Themaat, E. V. L. & Schulze-Lefert, P. *Proc. Natl Acad. Sci. USA* **111**, 585–592 (2014).
7. Edwards, J. *et al. Proc. Natl Acad. Sci. USA* **112**, E911–E920 (2015).
8. Horton, M. W. *et al. Nature Commun.* **5**, 5320 (2014).
9. Rappé, M. S. & Giovannoni, S. J. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).

This article was published online on 2 December 2015.

## CLIMATE SCIENCE

# A history of Greenland's ice loss

**Aerial photographs, remote-sensing observations and geological evidence together provide a reconstruction of mass loss from the Greenland Ice Sheet since 1900 — a great resource for climate scientists. [SEE LETTER P.396](#)**

BEATA M. CSATHO

**L**oss of ice-sheet mass is a major contributor to current sea-level rise, and is expected to continue as global warming proceeds<sup>1</sup>. Detailed reconstructions of changes in the Greenland and Antarctic ice sheets over the past few decades are available, based on remotely sensed data. But extending this record further into the past poses a

big problem because of the lack of systematic monitoring of changes in ice-sheet elevations. On page 396 of this issue, Kjeldsen *et al.*<sup>2</sup> present the first observation-based estimate of mass loss from the Greenland Ice Sheet from the end of the nineteenth century, when it began to retreat from its maximum extent achieved during the Little Ice Age (LIA), to the present day. Their findings show how the reconstruction of past ice-sheet changes





**Figure 1 | The Upernavik Ice Stream in northwest Greenland.** This glacier is one of many that drain the Greenland Ice Sheet into the sea. A trimline — distinguished by differently coloured rock above and below the line — is visible in the nearby hill, and indicates the maximum extent of ice during the Little Ice Age (LIA). Kjeldsen *et al.*<sup>2</sup> report a reconstruction of mass loss from the Greenland Ice Sheet since 1900, the end of the LIA.

helps to account for sources of sea-level rise and improves our understanding of the major processes controlling ice-sheet mass loss.

Historical photographs of ice sheets provide long-term context for mass loss by enabling measurements of their surface elevations and extent before satellites were used for remote sensing. Moreover, they facilitate the accurate mapping of glacial geomorphic features such as vegetation trimlines (Fig. 1) and moraines, which mark the highest extent of the ice sheet during the LIA in the case of the Greenland Ice Sheet. A treasure trove of aerial photographs of Greenland has been extensively used for many years<sup>3</sup>, but because of the difficulty in obtaining accurate surface-elevation measurements from historical photographs, a detailed timeline of mass loss was reconstructed only for the largest glaciers<sup>4,5</sup>.

To estimate the mass-balance history of the Greenland Ice Sheet — the time course of differences between mass gained by snow accumulation and that lost by melting and calving of icebergs — since the LIA, Kjeldsen *et al.* began by reprocessing images taken by the comprehensive Greenland aerial photography survey during 1978–87. They used modern photogrammetric methods to derive high-resolution, accurate digital elevation models (DEMs) depicting the ice-sheet surface at the sheet's margins during the survey period. They also reconstructed the ice-sheet margins during the LIA in three dimensions by mapping vegetation trimlines and glacial moraines. Taken together with laser-altimetry measurements from 2003 to 2010, these analyses enabled the authors to determine elevation changes for three different epochs since the 1900s.

The results show that the Greenland Ice

Sheet contributed substantially to sea-level rise throughout the twentieth century, providing at least  $25 \pm 9.4$  millimetres of the total global mean rise. Furthermore, rates of mass loss during 2003–10 were twice those during the twentieth century, mostly because of increasing water runoff from the surface, whereas discharge through iceberg calving has remained essentially the same since the LIA.

Kjeldsen and colleagues also report a large spatial variation in ice-sheet changes, indicating that the sheet's response to climate forcing is modulated by local geometric factors such as the topography of the underlying bed and the sizes of the drainage basins of individual glaciers. The striking similarity between the elevation-change patterns during the different epochs suggests that local controls act similarly on both decadal and centennial timescales.

The authors' discovery of a large mass loss, which averaged 75 gigatonnes per year (equivalent to a sea-level rise of  $0.21 \text{ mm yr}^{-1}$ ) during the twentieth century, emphasizes the need for improvements to the record of ice-sheet changes before the start of detailed remote-sensing measurements in the 1990s. Existing long-term records are usually based on time series of the positions of ice-sheet margins, but such records can be misleading for glaciers that flow into the ocean, whose floating termini can advance or retreat without any substantial changes farther up-glacier. Furthermore, only repeated elevation measurements allow the quantification of mass loss that is necessary to estimate contributions to sea-level rise. Kjeldsen and co-workers' results provide an excellent framework for selecting regions that represent different long-term mass-loss patterns for further detailed studies.

A crucial objective of those studies should be to examine the stability of the Greenland Ice Sheet between 1960 and 1990. It has been assumed that this ice sheet was in equilibrium during this period, and so calculated changes in its surface mass-balance relative to the average during 1960–90 are used to work out whether recent ice-sheet surface losses are anomalous<sup>6</sup>. Kjeldsen *et al.* challenge this assumption by arguing that it contradicts the long-term persistent mass loss detected in their study. However, the temporal sampling of their study is not sufficiently detailed to rule out the possibility that a near-steady-state condition existed following the warm period that occurred in the 1930s and 1940s.

The rich archive of historical stereo aerial photographs of Greenland includes: systematic surveys taken during the 1930s that were used to generate 1:250,000 scale topographic maps; oblique aerial photographs taken by the US Air Force for reconnaissance during the Second World War using a Trimetrogon camera (which also enable topographic information to be determined); and repeat surveys of the catchment basins of all major outlet glaciers around the Jakobshavn Isbræ during 1957–58 and in 1964, taken as part of the International Glaciological Expeditions to Greenland. Moreover, high-resolution stereo images collected by US intelligence satellites are available from the 1960s and 1970s. If all of these were combined with more-recent satellite observations, then a comprehensive record of long-term surface elevation, positions of calving fronts and ice margins, and ice-velocity changes could be obtained. This could be used to assess the implications of recent changes in the context of climate change and to provide input for modelling studies.

In the meantime, the authors' reconstruction will help to improve numerical models — by providing a time series of changes at ice sheet margins for the whole Greenland Ice Sheet during the twentieth century, suitable for validating models. Although the extensive spatial overlap of laser altimetry and DEMs derived from stereo photogrammetry along the ice-sheet margins provides robust and accurate change detection in these regions between the 1980s and the present, further research — particularly the use of more-realistic ice-sheet models — is needed to derive accurate elevations within the interior of the ice sheet before the start of laser-altimetry observations in the 1990s. Improving the accuracy of past elevation reconstructions would result in better estimates of long-term mass-balance changes.

Finally, once the timing of equilibrium conditions for the Greenland Ice Sheet is verified, a detailed reconstruction for that period could serve as a steady-state ice-sheet surface for initializing ice-sheet models. Establishing such a steady-state surface is a prerequisite for deriving projections of future ice-sheet evolution

that are more credible than currently available projections. ■

**Beata M. Csatho** is in the Department of Geological Sciences, University at Buffalo, New York 14260, USA.  
e-mail: [bcsatho@buffalo.edu](mailto:bcsatho@buffalo.edu)

1. Church, J. A. et al. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the*

*Intergovernmental Panel on Climate Change* (eds Stocker, T. F. et al.) 1137–1177 (Cambridge Univ. Press; 2013).

2. Kjeldsen, K. K. et al. *Nature* **528**, 396–400 (2015).
3. Weidick, A. *Grønlands Geol. Undersøgelse Bull.* **73** (1968).
4. Csatho, B., Schenk, T., van der Veen, C. J. & Krabill, W. B. *J. Glaciol.* **54**, 131–144 (2008).
5. Khan, S. A. et al. *Cryosphere* **8**, 1497–1507 (2014).
6. van den Broeke, M. et al. *Science* **326**, 984–986 (2009).

## In retrospect

# Twenty-five years of the sex-determining gene

**The discovery that the gene *SRY* on the mammalian Y chromosome drives testis development marked a turning point in the decades-long quest to understand the genetic underpinnings and evolution of sex determination.**

JENNIFER A. MARSHALL GRAVES

It has long been known that a testis-determining factor (TDF) on the Y chromosome kick-starts testis development in humans and other mammals. The testes make hormones, and these hormones make the embryo male. Twenty-five years ago, Sinclair *et al.*<sup>1</sup> reported in *Nature* that TDF was the gene *SRY*. This discovery opened up the surprisingly intricate genetic pathway that determines whether a baby is born a boy or a girl. It also led to an understanding of how genes on the Y chromosome evolved, and of the impact of this key evolutionary event.

Until the 1980s, there was no viable candidate sex-determining gene. Just where was TDF located? What kind of product did it encode? What did it do? During the 1980s, the position of TDF was narrowed down to a small region on the short arm of the Y chromosome, when it was found that some males had XX chromosomes that harboured a small piece of the Y, whereas some females had XY chromosomes that lacked bits of the Y — these added and deleted regions of Y were assumed to contain the TDF sequence. The race was then on to find TDF.

In 1987, the geneticist David Page and his associates<sup>2</sup> identified the first coding gene on the human Y, called *ZFY*. The gene looked like a winning candidate: it was in the right place; it was expressed in the testis; and it was evolutionarily conserved in other placental mammals, such as monkeys, mice, dogs and horses. But in 1988, PhD students in my laboratory<sup>3</sup>, Andrew Sinclair and Jamie Foster, mapped *ZFY* to a non-sex chromosome (an autosome) in marsupials, which are a separate branch of mammals. A few months later, it

was found<sup>4</sup> that, although *ZFY* is expressed in mouse sperm precursors, it is absent from the other cells of the testis, where a true TDF must be expressed to exert a sex-determining effect.

Sinclair joined a renewed hunt for human TDF in the laboratory of geneticist Peter Goodfellow, using DNA from XY males that had even smaller pieces of the Y than had previously been studied. This was slow and frustrating work, because the Y chromosome is full of repetitive sequences and so specific regions are hard to pinpoint. It was 1990 before they found<sup>1</sup> a small coding gene close to the end of the Y chromosome (Fig. 1). Noncommittally they called the gene *SRY*, for sex region on the Y. The final proof that *SRY* was the TDF came from the discovery of *SRY* mutations in XY females<sup>5</sup> and from the demonstration that adding *Sry* to XX mice was sufficient to induce male development<sup>6</sup>. *SRY* was located on the Y in other placental mammals and, thankfully, even in marsupials<sup>7</sup>.

Researchers in the field imagined that identifying TDF would rapidly lead to an understanding of how it worked, and would point to other genes in the sex-determining pathway. But 25 years on, it has become clear that the pathway kick-started by *SRY* is complex, full of checks and balances.

Initially, *SRY* proved a puzzle because it was unlike any known gene. It turned out to be a member of a previously unidentified family, now called the SOX genes. Painstaking biochemical studies of the *SRY* protein revealed that it bound to a certain DNA sequence and bent it at an angle, presumably to bring other sequences — or the proteins bound to them — into proximity, promoting or inhibiting transcription<sup>8</sup>. The discovery of a different



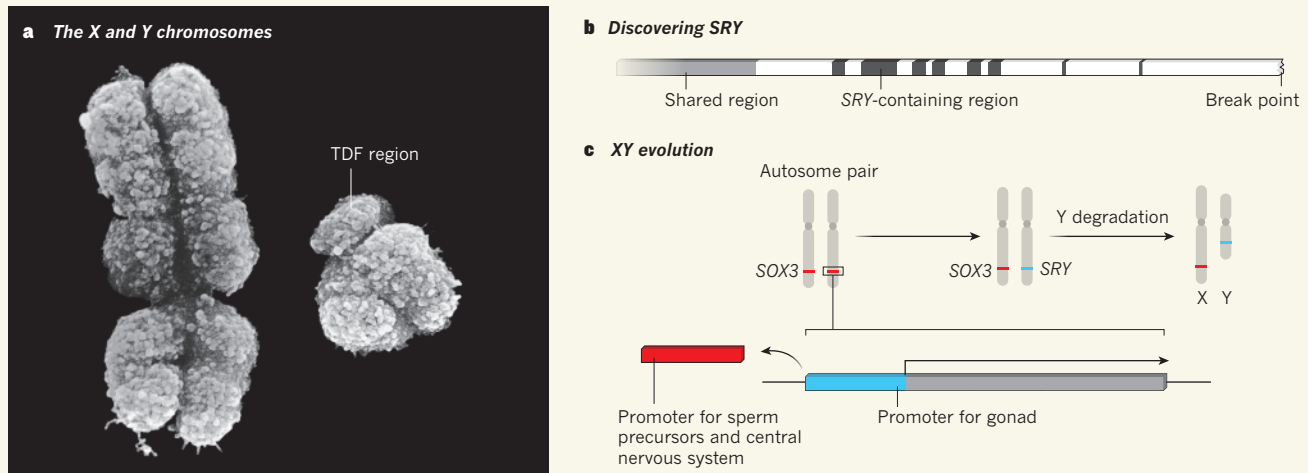
## 50 Years Ago

**The Royal Society Anniversary Address** by Lord Florey, O.M., P.R.S. Perhaps the deployment of Government resources is the modern equivalent of events in the early days of the Society when Fellows contributed—or sometimes did not contribute—a shilling a week towards demonstrating experiments at meetings. There never was enough money... At the moment it is considered to be desirable to give free medicine to all. The application of free calamine lotion to the irritated skins of the populace may be more important than administering to the needs of irritated scientists; but this sort of judgement is in the realm of politics... it has long been the policy of the Society to have symposia and lectures... the popularity of such gatherings has brought difficulties... on one occasion, we had to migrate to the lecture theatre of the Shell Building on the South Bank... one consequence of this peripatetic existence has been that we have had to procure a coffin-like box for the transport of the mace, and I am sure that our original Fellows, and even Charles II himself, might have been somewhat astonished at the adventures of their royal emblem.  
**From *Nature* 18 December 1965**

## 100 Years Ago

The Romanes Lecture... was a scathing indictment of the ineptitude of the lawyer-politicians who possess a dominating influence on national affairs... To the neglect of science, and the excessive predominance in Parliament and the Government of men with the spirit of the advocate to whom all evidence which will not support their case is unwelcome, Prof. Poulton ascribes the chief mistakes in the conduct of the war.  
**From *Nature* 22 November 1915**





**Figure 1 | An evolving understanding of sex.** **a**, In humans, sex is based on the presence or absence of the Y chromosome, seen here with its larger partner, X. The testis-determining factor (TDF) that drives male development was known to lie on the short arm of Y, but its identity was a mystery. **b**, In 1990, Sinclair *et al.*<sup>1</sup> found two males with only a small piece of Y, which had been broken and fused to the X. They scoured the 35,000 base pairs between the break points and the region at the tip of the Y that is shared with the X, finding several regions (black) that were specific to the Y. One of these regions contained the TDF gene, SRY. **c**, This discovery led to

an understanding of how X and Y evolved. The gene SOX3 was located on a pair of non-sex chromosomes (autosomes) in the ancestors of mammals. A promoter sequence drove expression of SOX3 in sperm precursors and the central nervous system. The promoter on one copy of SOX3 was replaced with a sequence that drives expression in the undifferentiated gonad (a tissue that can develop into either an ovary or a testis). This expression pattern allowed the new gene, SRY, to direct testis development. Over time, genes not needed for male development were degraded on this chromosome, giving rise to the Y. (Part **b** adapted from ref. 1.)

SOX gene that was disrupted in XY female babies with a severe bone deformity<sup>9,10</sup> revealed that this gene, SOX9, is the binding target of SRY protein. SOX9 is now known to be a master regulator of sex determination throughout the vertebrates.

Studying the mutations that cause sex reversal in humans, mice, goats or dogs (the same pathway is active in all mammals) has proved a successful strategy for identifying many genes in the sex-determination pathway. Gradually, a network of genes that are regulated by, or regulate, SRY or SOX9 has been constructed, and their function tested by mutating the genes in mice<sup>11</sup>. Some genes promote testis formation, some maintain it, and yet others oppose them. This pathway and its control is still being explored. Our improved understanding has helped us both to answer fundamental scientific questions and to diagnose and treat many babies who are born with disorders of sex determination<sup>12</sup>.

The other major line of research enabled by the identification of SRY was the evolution of sex genes and chromosomes. The hunt for SRY in marsupials revealed that mammals have an SRY-related gene on the X chromosome, SOX3, which was proposed to be the ancestor of SRY<sup>13</sup>. This idea is supported by human and mouse data<sup>14</sup> that showed that misexpression of SOX3 in the undifferentiated gonad (a tissue can develop into either an ovary or a testis, depending on the signals it receives) drives male development in XX embryos. SRY probably evolved from SOX3 when its 5' region was replaced by a promoter sequence that drove expression in the gonad (Fig. 1).

Although it might seem counterintuitive that the testis-determining factor evolved from the X chromosome, it has since emerged<sup>15</sup> that 20 of the 27 genes on the male-specific part of the human Y evolved from genes on the X. Thus, the Y is basically a degraded X chromosome. This supports the hypothesis that sex chromosomes originate when one member of an autosome pair acquires a sex-determining gene. Nearby genes then also acquire a sex-specific function, crossing over between the chromosome pair is suppressed to keep the male-specific gene package together, and the genetically isolated region on the sex-specific chromosome degrades rapidly<sup>15,16</sup>.

**“It has become clear that the pathway kick-started by SRY is complex, full of checks and balances.”**

The mammalian XY sex pair was probably defined by the evolution of SRY. Vertebrate phylogeny puts the age of SRY and the XY pair at between 166 million and 190 million years old. Furthermore, rapid speciation in other lineages that have undergone sex-chromosome turnover raises the possibility that acquisition of SRY might have driven the divergence of the egg-laying monotreme mammals from the rest of the mammalian lineage — monotremes have a bizarre, complex sex-determination system that is related to bird sex chromosomes<sup>17</sup>.

The future of the Y chromosome is now hotly debated. Evidence suggests that the mammalian Y will disappear in just a few million years if gene loss continues at the same rate as in the past<sup>18</sup>. It has already disappeared

in two groups of rodents, and SRY has been replaced by another gene from the sex-determining network<sup>19</sup>. The primate Y seems more stable<sup>20</sup>, but will eventually erode away. Humans may be in for another round of sex-chromosome turnover — and maybe speciation — if and when SRY finally bows out. ■

**Jennifer A. Marshall Graves** is at the School of Life Science, La Trobe University, Melbourne, Victoria 3086, Australia, and at the Research School of Biology, Australian National University, Canberra. e-mail: j.graves@latrobe.edu.au

1. Sinclair, A. H. *et al.* *Nature* **346**, 240–244 (1990).
2. Page, D. C. *et al.* *Cell* **51**, 1091–1104 (1987).
3. Sinclair, A. H. *et al.* *Nature* **336**, 780–783 (1988).
4. Koopman, P., Gubbay, J., Collignon, J. & Lovell-Badge, R. *Nature* **342**, 940–942 (1989).
5. Berta, P. *et al.* *Nature* **348**, 448–450 (1990).
6. Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. & Lovell-Badge, R. *Nature* **351**, 117–121 (1991).
7. Foster, J. W. *et al.* *Nature* **359**, 531–533 (1992).
8. Harley, V. R. & Goodfellow, P. N. *Mol. Reprod. Dev.* **39**, 184–193 (1994).
9. Foster, J. W. *et al.* *Nature* **372**, 525–530 (1994).
10. Wagner, T. *et al.* *Cell* **79**, 1111–1120 (1994).
11. Eggers, S., Ohnesorg, T. & Sinclair, A. H. *Nature Rev. Endocrinol.* **10**, 673–683 (2014).
12. Ohnesorg, T., Vilain, E. & Sinclair, A. H. *Sex. Dev.* **8**, 262–272 (2014).
13. Foster, J. W. & Graves, J. A. M. *Proc. Natl Acad. Sci. USA* **91**, 1927–1931 (1994).
14. Sutton, E. *et al.* *J. Clin. Invest.* **121**, 328–341 (2011).
15. Graves, J. A. M. *Cell* **124**, 901–914 (2006).
16. Charlesworth, B. *Science* **251**, 1030–1033 (1991).
17. Veyrunes, F. *et al.* *Genome Res.* **18**, 965–973 (2008).
18. Aitken, R. J. & Graves, J. A. M. *Nature* **415**, 963 (2002).
19. Kuroiwa, A. *et al.* *Chromosome Res.* **19**, 635–644 (2011).
20. Hughes, J. F. *et al.* *Nature* **437**, 100–103 (2005).

# Rarity in mass extinctions and the future of ecosystems

Pincelli M. Hull<sup>1</sup>, Simon A. F. Darroch<sup>2,3</sup> & Douglas H. Erwin<sup>2</sup>

**The fossil record provides striking case studies of biodiversity loss and global ecosystem upheaval. Because of this, many studies have sought to assess the magnitude of the current biodiversity crisis relative to past crises—a task greatly complicated by the need to extrapolate extinction rates. Here we challenge this approach by showing that the rarity of previously abundant taxa may be more important than extinction in the cascade of events leading to global changes in the biosphere. Mass rarity may provide the most robust measure of our current biodiversity crisis relative to those past, and new insights into the dynamics of mass extinction.**

It has become commonplace to refer to the modern biodiversity crisis as the ‘sixth mass extinction’<sup>1,2</sup>. With three short words, we place the biotic and environmental disturbance created by mankind on par with the greatest biodiversity crises of the past half billion years. This is a comparison that demands close attention as the ‘Big Five’ mass extinctions include truly catastrophic events<sup>3,4</sup>, the biggest of which resulted in the inferred extinction of >75% of species alive at the time<sup>1,4</sup>. In addition, mass extinctions have shaped the evolutionary history of the planet<sup>5–7</sup>. Organisms that were ecologically dominant before a mass extinction frequently do not survive, and rarely enjoy the same levels of dominance in the aftermath<sup>6,8</sup>. However, there are fundamental differences between the types of data upon which past mass extinctions have been identified, and those upon which the current biodiversity crisis is being assessed. That is, abundant marine fossil genera on multi-million year timescales for the former<sup>9,10</sup>, and (often rare) terrestrial species on decadal to centennial timescales for the latter<sup>1</sup>. So the question is critical: are we currently in the midst of the ‘sixth’ mass extinction, and can we develop an appropriate metric for the comparison of ancient and modern biotic crises?

The Big Five mass extinctions were profoundly disruptive events with effects extending far beyond the loss of taxonomic diversity<sup>11–15</sup>. In addition to extinction, all major mass extinctions are also characterized by prolonged intervals of ecological change<sup>12,16</sup>. Ecosystems are comprised of interacting networks of biotic and biophysical components, including taxa, nutrients, and their trophic and non-trophic interactions<sup>17</sup>. Species loss and ecosystem reassembly during mass extinction is unsurprising given the disruption of ecological networks<sup>18</sup>. For hundreds of thousands to millions of years after mass extinctions, a series of short-lived, low-diversity and (at times) low productivity ecosystems dominate<sup>16,19,20</sup>. Large-bodied taxa often become dwarfed, or are replaced by small-bodied taxa<sup>21,22</sup>. Previously dominant groups may be supplanted in the evolutionary diversifications that follow<sup>23–25</sup>, as new, diverse ecosystems are built<sup>26</sup>. The largest extinction intervals result in permanent state changes in the structure of ecosystems, as well as the character of the flora and fauna that dominate them<sup>5,25,27</sup>. Mass extinctions, therefore, not only punctuate the history of life, they also forever alter its trajectory.

In this light, the fossil record of mass extinctions is an important laboratory for understanding the effects of current environmental change on global ecosystem structure and function<sup>28</sup>. A key question is: how do minor biodiversity crises become mass extinctions? And, why do mass extinctions tend to coincide with permanent state changes in global ecosystems? To date, studies have considered these issues by comparing

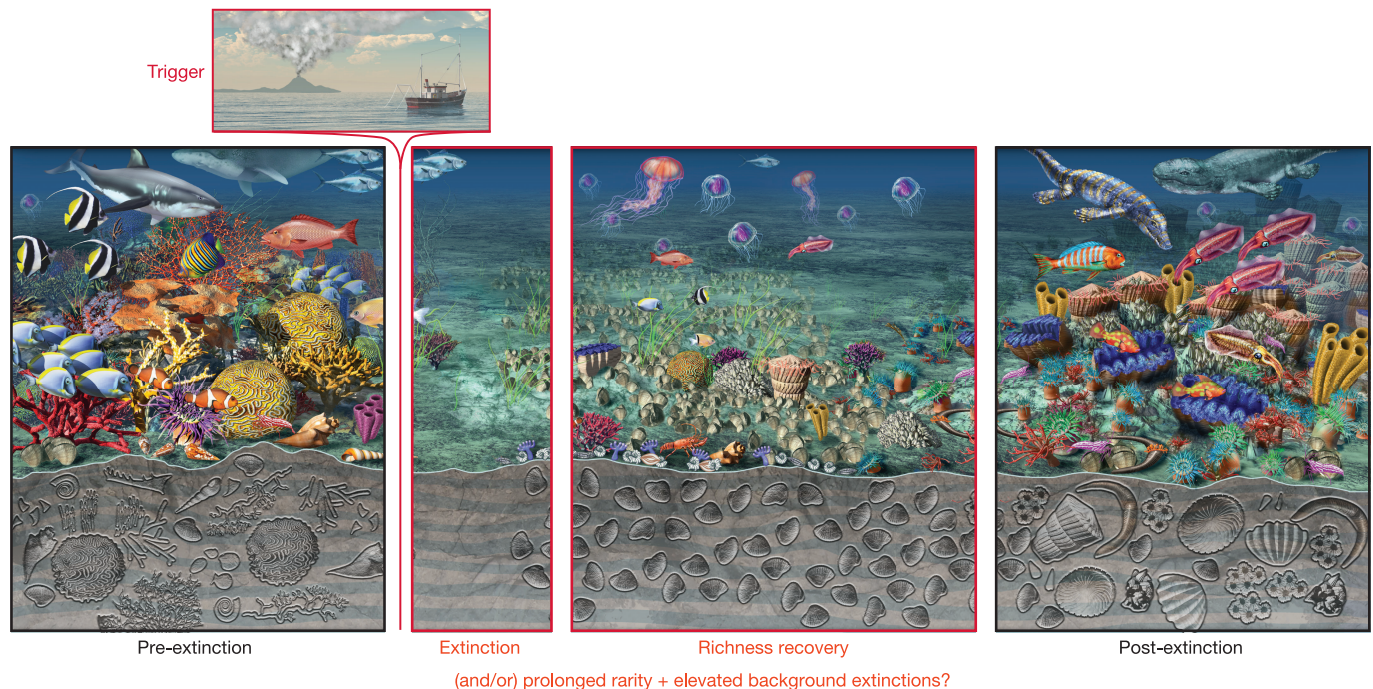
projected rates of modern species loss and rates estimated from the fossil record<sup>1,11,29</sup>—a method complicated by the need to extrapolate across temporal scales and abrupt state changes. Here, we propose a different approach, and consider whether the loss of species abundance—mass rarity—might have characterized past mass extinctions as they were occurring. Rarity is important for two reasons: first, because it more accurately reflects function in ecological networks<sup>30</sup> and thus mass rarity (rather than mass extinction) may be a primary driver of the events and patterns associated with the mass disappearance of fossils from the fossil record. Second, the extent to which previously common taxa have become rare offers a direct metric of the size of the present biotic crisis. There may be no need to project current extinction rates in order to get a sense of the future of ecosystems. Mass rarity may be all that is needed to forever change the biosphere.

## From past abundance to current rarity

Humans have reduced the abundance of many historically common species. This increased rarity has been achieved through wholesale reduction in geographic ranges and/or population sizes, through modification of terrestrial habitats, appropriation of primary productivity for humanity, overexploitation and pollution, among other factors<sup>31–33</sup>. On land, widespread evidence exists for ongoing habitat loss and population declines globally<sup>31,34</sup>. This includes, for instance, a 20% decline in habitat specialist populations monitored by the Wild Bird Index since the 1980s, and continuing declines in the IUCN Red List Index of species survival aggregated across birds, mammals, amphibians and corals<sup>31</sup>. Likewise, most fished coral reefs support less than half the expected fish biomass<sup>35</sup>, with long-term declines in the abundance of reef taxa since first human contact<sup>36</sup>. Among subsets of mammals, birds, butterflies, and highly mobile pelagic predators, more than 50% of the taxa studied have experienced range contractions in the last decades to centuries<sup>37–39</sup>. Yet to date, the absolute number of recorded species extinctions is dwarfed by those inferred for mass extinctions in the geological past<sup>1,11</sup> and local declines in species richness are equivocal<sup>33,40</sup>. However, the extent of abundance loss is not equivocal, nor is the effect of land use<sup>34</sup>. Mass rarity, that is the reduction in geographic range and/or numerical abundance of a species globally, seems to be one or more orders of magnitude more severe than extinctions to date<sup>41–44</sup>, and is an urgent conservation priority for both species and ecosystems<sup>38,45–47</sup>. What remains a major unknown, however, is how global mass rarity today relates to the biotic crisis recorded in the fossil record, and what sustained mass rarity might mean for the future of ecosystems.

<sup>1</sup>Department of Geology and Geophysics, Yale University, New Haven, Connecticut 06520-8109, USA. <sup>2</sup>Department of Paleobiology, National Museum of Natural History, Washington, DC 20013-7012, USA. <sup>3</sup>Department of Earth and Environmental Sciences, Vanderbilt University, Nashville, Tennessee 37235-1805, USA.





**Figure 1 | Mass rarity and mass extinction are indistinguishable in the fossil record, and may have the same ecosystem effects.** Anthropogenic activities have led to mass rarity of many previously abundant flora and fauna (right to middle). Mass rarity can look like mass extinction in the fossil record because the previously abundant taxa become so rare as to no longer be readily observed (bottom). Previously abundant and ecologically important

groups, such as ecosystem engineers may not actually become extinct, but decline below the abundance threshold required for them to perform their ecological roles, becoming ecological ‘ghosts’. Chance reassembly after mass rarity could lead to drastically different ecosystem structure and function even with minimal extinction (right)—raising the question of what the future might hold. Artwork courtesy of Nicolle R. Fuller, Sayo-Art.

We suggest that global rarity today (that is recent mass rarity, not the local rarity of most species in ecological studies as in ref. 48) may already be equivalent to intervals of pervasive fossil disappearance (Fig. 1). This is because the fossil record, particularly as it is preserved and studied across extinction boundaries (Box 1), primarily records the dynamics of durably skeletonized, geographically widespread, abundant taxa, and not the absolute presence or absence of all species originally in that ecosystem. When taxa are rare they can be missed, and when events are rapid, the order and importance of different factors can be hard to interpret.

The vast majority of species evolve, exist and become extinct without being preserved as fossils<sup>49–51</sup>. The fossil record is instead dominated by species that inhabit environments with high preservation potential. Such environments include those in which sediment accumulates, such as in (or around) lakes, rivers, swamps, marine basins, or reef tracts<sup>52</sup>. Even in such areas, most species stand little chance of being preserved. Rather, the fossil record is dominated by those taxa possessing heavily mineralized hard parts, such as teeth, bone or shells<sup>51</sup>. Organisms that are very small, entirely soft-bodied, or occur in ephemeral habitats are rarely preserved<sup>49–51</sup>. Additionally, as in living ecosystems, species that exist over a broad geographic range and in large numbers have a higher probability of being found than species that are rare and/or geographically restricted.

As a consequence, the fossil record of abundant, widespread, hard-bodied, marine taxa shapes our paleontological perspective of the long-term dynamics of life<sup>10</sup> (see Box 1). By definition, a mass extinction is an interval of time characterized by elevated rates of extinction relative to background intervals<sup>14,15</sup>. In practice, however, they are identified by the geologically sudden disappearance of abundant, long-lived genera (or higher order taxa) from global-scale compilations of fossil occurrences of biomineralizing taxa<sup>9,10</sup>.

The often-discussed ‘Big Five’ mass extinction events were first recognized in this way from the shelly marine fossil record: the end Ordovician (~445 million years ago (Ma)), end Devonian (~375 Ma), Permo–Triassic (PT; 251 Ma), Triassic–Jurassic (TJ; 199 Ma), and Cretaceous–Palaeogene

(KPg; 66 Ma)<sup>10,15</sup>, although marine and terrestrial extinctions have subsequently been shown to often go hand-in-hand<sup>53</sup>.

Detecting and predicting the ultimate severity of a mass extinction as it is happening requires a detailed understanding of the triggers and feedbacks of the extinction interval—the geologically brief interval of time when previously abundant fossil taxa disappear en masse (see Extinction in Fig. 2). Assessments of the severity of the current biodiversity crisis relative to those of the past presuppose an understanding of these geologically near-instantaneous events (Box 1). So, how much is actually known?

## Changing the world

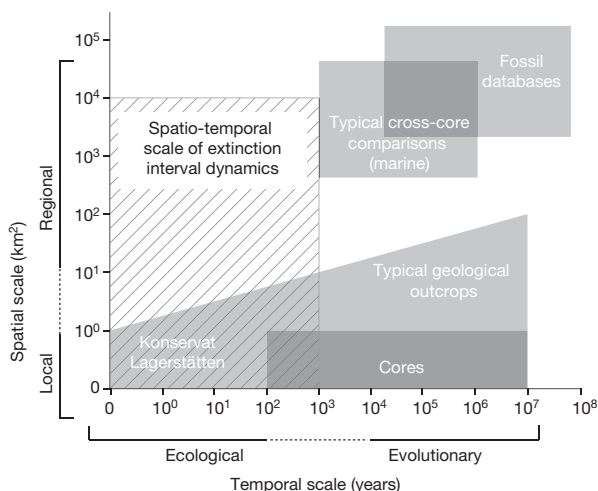
Extinction intervals involve a primary trigger, secondary feedbacks, ecological transitions, and extinction (Fig. 2)<sup>18</sup>. The primary trigger (or set of triggers) is the environmental disturbance(s) that precipitates the mass extinction—including, for instance, asteroid impact or massive volcanism<sup>12</sup>. A primary trigger need not drive many species extinct, as per the classic view of mass extinctions (Fig. 3a, scenario 1). Rather, it need only cause sufficient disturbance for processes like extinction debt<sup>54,55</sup> or ecological collapse<sup>18</sup> to result in mass secondary extinctions (Fig. 3b, scenario 2). A primary trigger might produce widespread rarity of formerly dominant taxa, thereby greatly elevating rates of background extinction for these taxa (Fig. 3c, scenario 3), or could directly cause the extinction of all species lost in a given interval. In addition, ecological turnover may precede the loss of taxa (that is, be driven by the primary trigger) or follow it (that is, result from the loss of species during extinction).

The brevity of mass extinctions (Box 1), combined with the time-averaged nature of the fossil record, currently precludes an understanding of the relative contribution of these four processes (Fig. 3). This makes it very difficult to use fossil data to disentangle alternative scenarios of the dynamics of mass extinctions: ‘trigger kills all’ (Fig. 3a), ‘trigger sparks feedbacks and secondary extinctions’ (Fig. 3b), and ‘trigger drives mass rarity and elevated extinction risk’ (Fig. 3c). We have little information yet about the relative importance of primary and secondary extinctions or mass rarity during past events.

## BOX 1

## The scale of extinction dynamics

Extinction intervals are extremely short (Fig. 2), even geologically instantaneous, relative to the typical resolving power of the fossil record<sup>112</sup> (see Box Figure). The three mass extinction events with the best geochronologic constraints on their duration (PT, TJ and KPg) all occurred on time scales on the order of  $10^3$ – $10^4$  years<sup>18,113–115</sup>. In exceptional circumstances, rapid sedimentation may preserve a temporally detailed record of a mass extinction in a local region<sup>114</sup>. However, taphonomic and sedimentological processes typically time-average accumulations of shell material such that individual samples will represent communities mixed over  $10^3$ – $10^4$  year intervals. We consider events ‘geologically instantaneous’ if they occur on timescales at or below the resolution of the records used to study them (here  $10^3$ – $10^4$  years). While exceptional ‘snapshots’ of the fossil seafloor during a single moment of time do exist (that is, Konservat Lagerstätten), they are so infrequent that they rarely figure in studies of mass extinctions, and none have yet been discovered crossing a major extinction boundary. The global paleontological and marine core compilations that are so key for detailing the broader patterns of extinction, currently lack the temporal resolution needed to disentangle the dynamics within the extinction interval. The unavoidable conclusion is that the ‘pixel size’ of the fossil record may be too temporally coarse, or spatially restricted, to resolve the most important processes during the extinction phase.



**Box Figure Mismatch in the spatio-temporal scale of ecosystems collapse and the resolving power of the fossil record.** The fossil record provides detailed records of macroevolutionary processes occurring at many spatial and temporal scales (shaded regions). The dynamics of extinction intervals occur on spatial and temporal scales just beyond those that are readily documented (striped box).

To be clear, these three scenarios are distinguished by the internal dynamics of the extinction interval (Figs 1 and 3). In scenario 1, the extinction of well-fossilized taxa is driven by the trigger and coincides with, or even precedes, major environmental change. In scenarios 2 and 3, mass extinction is delayed—being driven by secondary feedbacks or elevated background extinction risk, respectively—after profound ecological disruption.

Comparing the present crisis to those that have occurred in the past requires knowing which of these scenarios is typical or dominant, as each

involves distinct patterns of feedback, propagation of risk, and timing of extinction. To date, palaeontologists have acted on the implicit assumption that the first scenario is correct (with rare exceptions, as in refs 18, 56, 57), when all the fossil record indicates—at a minimum—is that there must have been a geological instantaneous loss in the abundance of previously dominant taxa at the extinction boundary (the third scenario). The relative importance of these scenarios during the extinction interval cannot be disentangled by standard quantitative paleontological approaches, like those used to estimate species ranges or to control for uneven sampling in diversity dynamics<sup>58</sup>, because the timescale of the extinction interval is much shorter than the uncertainty intervals associated with these approaches.

That said, the dynamics of modern ecosystems support the inference that mass rarity can drive permanent ecosystem change. Taxa need not go locally or globally extinct to destroy the links in an ecological network. Rather, species simply have to become so rare as to be ecologically insignificant<sup>59,60</sup>. For instance, in the Chesapeake Bay changes in land use (runoff, sedimentation and nitrification) and overfishing of oysters in the 19th and 20th centuries contributed to shift from a highly productive estuarine ecosystem with thriving oyster, crab and fish fisheries, to a eutrophic, oxygen-depleted, bacterially dominated system<sup>61,62</sup>. Likewise, overfishing of North Atlantic cod similarly resulted in a shift from a fish (cod)-dominated system to one dominated by invertebrates (shrimps, crab and lobster<sup>59,63</sup>). In the Caribbean, coral reefs collapsed after centuries of overfishing and pollution compounded by warming, coral bleaching, disease and invasive species, with widespread replacement of corals by macroalgae<sup>36,61,64</sup>. In each case, the new structure seems to be an alternative stable state, as extensive management efforts have been unable to restore historic ecosystem structure<sup>60,65</sup>.

The fossil record likewise documents examples of profound ecosystem change owing to shifts in the relative abundance (not just presence or absence) of taxa, including many of the turnovers in dominant reef builders<sup>66,67</sup>, the rise of angiosperms<sup>68</sup> and C4-grasses<sup>69</sup>, and during past biodiversity crises (see discussion below). In short, there is no a priori reason to believe that the extirpation of species drives observed ecosystem changes at mass extinction boundaries—global mass rarity may be as plausible a mechanism for ecosystem change as mass extinction. This being the case, we suggest that the extent of mass rarity might be the best metric for comparing the current crisis to those in the fossil record.

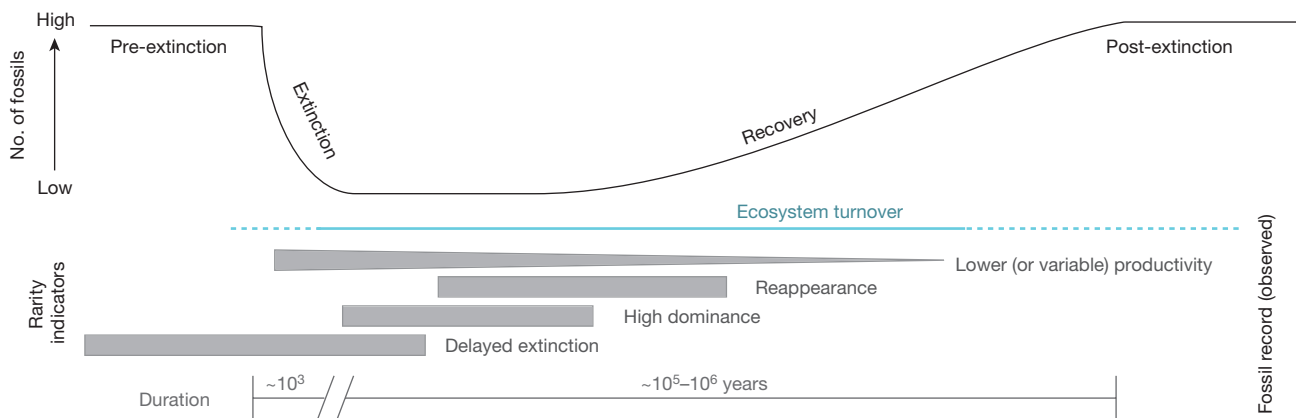
### The kill mechanism need only make the common rare

Although palaeontologists have focused on extinction more than rarity, they have identified unusual phenomena associated with rarity during mass extinction episodes. Rarity matters because geographically or numerically restricted taxa typically have a relatively small probability of being preserved in the fossil record, or being recovered by palaeontologists<sup>70</sup>. A species that undergoes a drastic reduction in population size, or contraction in range size, can thus appear to be ‘extinct’ in the fossil record, until that population either recovers, or eventually dies out entirely<sup>71,72</sup>.

Species that disappear from the fossil record—sometimes repeatedly, and often for millions of years—only to subsequently reappear are called ‘Lazarus’ taxa<sup>72</sup>. Such taxa are known from each of the Big Five mass extinctions boundaries<sup>72</sup>. They include a variety of clades with high preservation potential, such as molluscs across the PT extinction<sup>73</sup>, brachiopods across the Ordovician–Silurian<sup>74</sup> and KPg<sup>75</sup> extinctions, and ostracods across the late Devonian extinction<sup>76</sup>. Outside of extinction boundaries, once-abundant taxa can also vanish from the fossil record for  $10^5$ – $10^6$  years without extinction, owing to rarity. Striking examples include the coelacanth fishes (currently extant; ~70 million year fossil gap<sup>77</sup>) and the once widely abundant marine algae *Cyclagelosphaera* (currently extant; 54 million year fossil gap<sup>78</sup>).

Another example of extinction-related rarity is found in species that persist in low numbers through an extinction interval before dying out in the aftermath—a phenomenon known as ‘Dead Clades Walking’<sup>79,80</sup>.



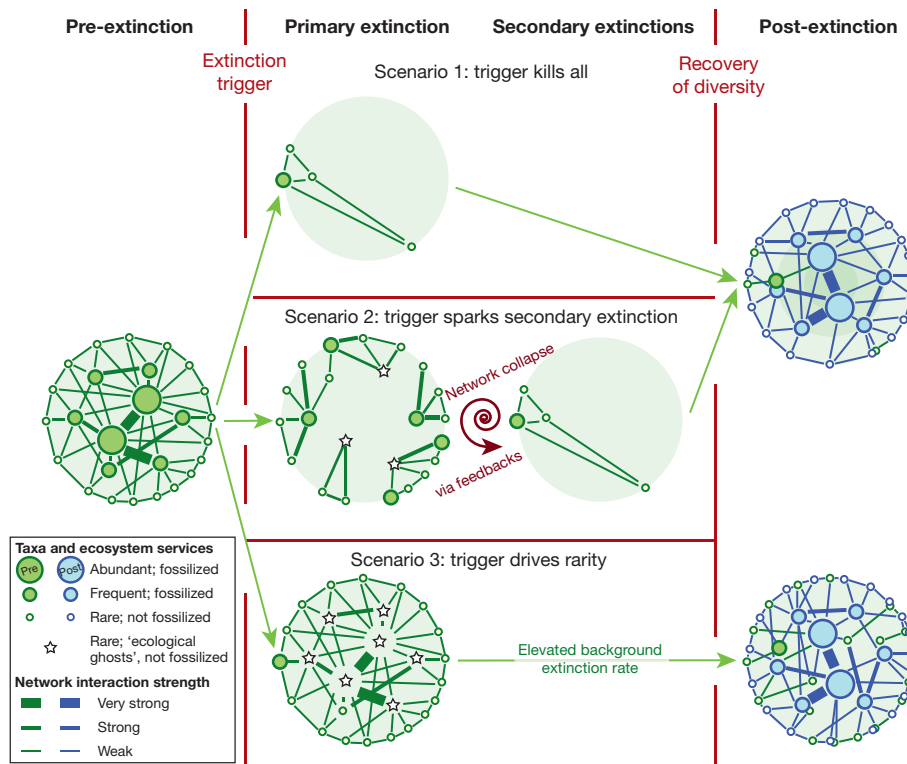


**Figure 2 | The sequence of taxonomic and ecosystem events across extinctions is unclear.** Extinction intervals have four recognized phases (at the top: pre-extinction, extinction, recovery, post-extinction), based on the richness of fossils preserved. The relationship between fossil diversity and changes in ecosystem structure and function is unclear and

may precede, coincide with, or follow the lost fossil diversity (blue solid to dashed line). A wide variety of palaeontological phenomena (grey boxes) document pervasive rarity as a feature of past mass extinctions. Most are widely accepted phenomena, with only the evidence for lowered productivity still debated within and among events<sup>56,87–91</sup>.

A frequently cited case is that of bellerophonitid gastropods after the PT extinction<sup>81</sup>. More generally, an estimated 10–20% of the genera surviving extinction intervals die out before global biodiversity recovers<sup>79</sup>. For other taxa we might imagine that the sudden loss of fossils across a boundary is driven by extinction or by persistent rarity. For the second case, rarity and range contractions at extinction boundaries can be followed by eventual extinction, long disconnected from the last fossil occurrence.

Three final attributes of past mass extinctions support the hypothesis of pervasive mass rarity. These features include the short-lived dominance of post-extinction taxa, the rarity of previously widespread habitats, and evidence for decreased primary productivity in the wake of extinctions. Those species that dominate assemblages immediately after extinctions are known as ‘bloom taxa’<sup>16</sup>. They have been recognized from the major, as well as many minor, extinction events<sup>16,20,71,82,83</sup>. The ecological success of post-extinction dominants in the unusual ecosystems characterizing



**Figure 3 | The geological brevity of mass extinctions makes it difficult to discern the relative importance of various processes.** Mass extinction intervals are geologically instantaneous, making it difficult to measure the processes responsible for determining the size and ecological impact of any event. Three major extinction interval scenarios are (top) scenario 1: the primary extinction trigger directly kills off the pre-extinction taxa, with the size and impact of extinction determined by trigger; (middle) scenario 2: the extinction trigger kills key taxa (or environmental resources) with feedbacks leading to secondary extinctions; or (bottom)

scenario 3: the trigger makes many species rare, many of which go extinct, and when abundant populations recover, the ecosystem, by chance, is structured differently. In scenarios 2 and 3 the decreased abundance in key taxa is sufficient to diminish their ecological effect (they become ecological ghosts) and precipitates further ecosystem collapse through secondary extinction and feedbacks. Also note that the primary trigger can be called the ‘kill-mechanism’ and include multiple coincident disturbances.

extinction aftermaths coincides with the prolonged rarity of all other taxa<sup>16,83,84</sup>. At the same time, pre-extinction habitats themselves often become rare or altered, as revealed by changes in the composition, continuity and texture of common sedimentary rock types<sup>20,73,85</sup>. In addition, the rate of sediment accumulation is often much lower during and after the extinction interval (for example, prolonged low sedimentation after the PT<sup>86</sup>), a feature due at least in part to the low abundance of fossil-forming organisms (as for pelagic sediments after the KPg<sup>87</sup>). This, and other lines of evidence<sup>56,87–89</sup>, have been used to argue for some suppression of primary productivity in the aftermath of extinctions—although to what extent this is true is still hotly debated<sup>90,91</sup>. Regardless, these lines of evidence indicate that pervasive rarity of formerly abundant taxa is unifying feature of extinctions and their aftermaths.

This evidence for mass rarity during past extinction events is surprisingly similar to the widespread rarity of previously common flora and fauna today. The modern ocean is full of ecological ‘ghosts’—taxa that are so rare they no longer provide past ecological services<sup>36,38,92,93</sup>. Mass rarity includes local, often remarkable, declines in species abundance, as well as range contractions (as reviewed in refs 38 and 44). For those species with excellent historical and fossil records, like Caribbean corals, the recent population collapse contrasts with the marked resilience to past climatic perturbation<sup>36,94,95</sup>. What’s more, the loss of species abundance is known to, at times, have cascading effects on ecosystem structure and function<sup>45</sup>, and extinction debt may cause extinction hundreds<sup>96</sup> to millions<sup>97</sup> of years after an environmental perturbation. In this light, the paucity of extinctions in the oceans to date should not be viewed as a sign of the relative health of marine ecosystems<sup>11,38</sup>—rarity itself may be the most direct metric of how close global ecosystems are to a permanent state shift.

## Saving the fossil record of today

The effect of humanity is so pervasive<sup>32,36,93</sup> that we are leaving a globally recognizable mark in the rock record<sup>98,99</sup>. Some scientists are seeking to formally recognize this moment as the ‘Anthropocene’<sup>100,101</sup>—defining it as the epoch of human-dominated earth systems<sup>98,99</sup>. As we consider humanity’s effect on the biosphere, we must recognize that this history is still being written in stone and it remains ours to shape. Thus our hypothesis of past mass extinctions as mass rarity events offers a to-do list for avoiding the ecological aftermath of catastrophic and global biotic crises.

For ecologists and conservation biologists, we have argued that, on timescales comparable to those studied today, past mass extinction events may have been characterized by the geologically instantaneous mass rarity of previously abundant, widespread, well-preserved species. This argument is supported by the nature of the rock record, in which the observed presence or absence of a fossil species depends as much on its abundance as its existence. The rarity of previously common taxa is the only factor tied with certainty to the profound ecological change observed across extinction boundaries. And rarity alone may be enough to drive permanent shifts in the earth system—long before ‘rare’ turns into ‘extinct’. Because of this, we argue that changes in the abundance and ranges of previously common taxa provide an additional, potentially more accurate, metric of the severity of the current biotic crisis relative to those in the past than do extrapolated extinction rates.

To date, the majority of extinction studies have been biased towards terrestrial species and charismatic megafauna<sup>102,103</sup> and we know relatively little about changes in the abundance and ranges of the shelly marine invertebrates that would provide a direct link to mass extinctions in the fossil record<sup>104</sup>. Rarity of previously common taxa matters. In order to avoid a mass-extinction-like fossil record, we need to increase the population size and geographic range of once-abundant taxa and trophic groups (that is, reverse defaunation and defloration) and minimize the geographic extent of habitat destruction.

From custodians of deep time<sup>105</sup>, we need quantitative assessments of the fossil record of the present and future earth in order to accurately size up current biotic changes with the same filter through which we see the past. Equally important will be studies of the dynamics and resilience

of full ecological networks (not just trophic food webs) during massive perturbations. Spatially explicit models of the various extinction scenarios (Fig. 3) would likewise aid in distinguishing among the potential mechanisms at play during mass extinctions<sup>18</sup>. Ongoing efforts to build palaeontological data archives<sup>106</sup> and to collect finely resolved records from extinction boundaries<sup>19,90,91</sup> are likewise key, as they provide the means to globally test emergent predictions on relevant timescales and key processes, like geographic rarity, on others<sup>107,108</sup>. Finally, the fossil record offers numerous examples of ecosystem change with and without fossil extinctions<sup>109,110</sup>. How and why this occurs is a key question to address if we are to predict, and avoid, a state shift in the structure and function of our biosphere in the years to come<sup>110</sup>. Although extinctions are rare<sup>44</sup>, the ecological ghosts of oceans past already swim in emptied seas<sup>11,111</sup>.

Received 5 May; accepted 15 October 2015.

- Barnosky, A. D. *et al.* Has the Earth’s sixth mass extinction already arrived? *Nature* **471**, 51–57 (2011).  
**A powerful marshalling of the paleontological evidence for a 6th mass extinction, in a paper that sparked much subsequent discussion and research.**
- Kolbert, E. *The Sixth Extinction: an Unnatural History* 1–319 (Holt, 2014).
- Alvarez, L. W., Alvarez, W., Asaro, F. & Michel, H. V. Extraterrestrial cause for the Cretaceous-tertiary extinction. *Science* **208**, 1095–1108 (1980).
- Erwin, D. H. *Extinction: How Life on Earth Nearly Ended 250 Million Years Ago* (Princeton Univ. Press, 2006).
- Wagner, P. J., Kosnik, M. A. & Lidgard, S. Abundance distributions imply elevated complexity of post-Paleozoic marine ecosystems. *Science* **314**, 1289–1292 (2006).  
**A key example of the profound potential of mass extinctions to permanently shift the structure of ecosystems.**
- Sahney, S., Benton, M. J. & Ferry, P. A. Links between global taxonomic diversity, ecological diversity and the expansion of vertebrates on land. *Biol. Lett.* **6**, 544–547 (2010).
- Jablonski, D. Mass extinctions and macroevolution. *Paleobiology* **31**, 192–210 (2005).
- Brusatte, S. L., Benton, M. J., Ruta, M. & Lloyd, G. T. Superiority, competition, and opportunism in the evolutionary radiation of dinosaurs. *Science* **321**, 1485–1488 (2008).
- Alroy, J. Dynamics of origination and extinction in the marine fossil record. *Proc. Natl Acad. Sci. USA* **105** (Suppl. 1), 11536–11542 (2008).
- Raup, D. M. & Sepkoski, J. J. Jr. Mass extinctions in the marine fossil record. *Science* **215**, 1501–1503 (1982).
- Harnik, P. G. *et al.* Extinctions in ancient and modern seas. *Trends Ecol. Evol.* **27**, 608–617 (2012).
- Hull, P. M. & Darroch, S. A. F. in *Ecosystems Paleobiology and Geobiology. The Paleontological Society Papers* Vol. 19 (eds A. M. Bush, S. B. Pruss, & J. L. Payne) 115–156 (Geological Soc. America, 2013).
- Erwin, D. H. Lessons from the past: biotic recoveries from mass extinctions. *Proc. Natl Acad. Sci. USA* **98**, 5399–5403 (2001).
- Bambach, R. K. Phanerozoic biodiversity mass extinctions. *Annu. Rev. Earth Planet. Sci.* **34**, 127–155 (2006).
- Sepkoski, J. J. in *Patterns and Processes in the History of Life* (eds D. M. Raup & D. Jablonski) 277–295 (Springer-Verlag, 1986).
- Erwin, D. H. The end and the beginning: recoveries from mass extinctions. *Trends Ecol. Evol.* **13**, 344–349 (1998).
- Schmitz, O. J. *et al.* From individuals to ecosystem function: toward an integration of evolutionary and ecosystem ecology. *Ecology* **89**, 2436–2445 (2008).
- Erwin, D. H. Temporal acuity and the rate and dynamics of mass extinctions. *Proc. Natl Acad. Sci. USA* **111**, 3203–3204 (2014).
- Hull, P. M., Norris, R. D., Bralower, T. J. & Schueth, J. D. A role for chance in marine recovery from the end-Cretaceous extinction. *Nat. Geosci.* (2011).
- Chen, Z.-Q. & Benton, M. J. The timing and pattern of biotic recovery following the end-Permian mass extinction. *Nat. Geosci.* **5**, 375–383 (2012).
- Twitchett, R. J. The Lilliput effect in the aftermath of the end-Permian extinction event. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **252**, 132–144 (2007).
- Payne, J. L. Evolutionary dynamics of gastropod size across the end-Permian extinction and through the Triassic recovery interval. *Paleobiology* **31**, 269–290 (2005).
- Droser, M. L., Bottjer, D. J., Sheehan, P. M. & McGhee, G. R. Decoupling of taxonomic and ecologic severity of Phanerozoic marine mass extinctions. *Geology* **28**, 675–678 (2000).
- Wood, R. *Reef Evolution* (Oxford Univ. Press, 1999).
- Sepkoski, J. J. A factor analytic description of the Phanerozoic marine fossil record. *Paleobiology* **7**, 36–53 (1981).
- Solé, R. V., Saldaña, J., Montoya, J. M. & Erwin, D. H. Simple model of recovery dynamics after mass extinction. *J. Theor. Biol.* **267**, 193–200 (2010).
- Bambach, R. K., Knoll, A. H. & Sepkoski, J. J. Jr. Anatomical and ecological constraints on Phanerozoic animal diversity in the marine realm. *Proc. Natl Acad. Sci. USA* **99**, 6854–6859 (2002).



28. Sepkoski, J. J. Jr. Biodiversity: past, present, and future. *J. Paleol.* **71**, 533–539 (1997).
29. Ceballos, G. *et al.* Accelerated modern human-induced species losses: entering the sixth mass extinction. *Science Advances* **1**, e1400253 (2015).
30. Naeem, S., Duffy, J. E. & Zavaleta, E. The functions of biological diversity in an age of extinction. *Science* **336**, 1401–1406 (2012).  
**A review of the multifarious impacts that a change in ecosystem structure can have on ecosystem function.**
31. Tittensor, D. P. *et al.* A mid-term analysis of progress toward international biodiversity targets. *Science* **346**, 241–244 (2014).
32. Halpern, B. S. *et al.* A global map of human impact on marine ecosystems. *Science* **319**, 948–952 (2008).
33. McGill, B. J., Dornelas, M., Gotelli, N. J. & Magurran, A. E. Fifteen forms of biodiversity trend in the Anthropocene. *Trends Ecol. Evol.* **30**, 104–113 (2015).
34. Newbold, T. *et al.* Global effects of land use on local terrestrial biodiversity. *Nature* **520**, 45–50 (2015).
35. MacNeil, M. A. *et al.* Recovery potential of the world's coral reef fishes. *Nature* **520**, 341–344 (2015).
36. Pandolfi, J. M. *et al.* Global trajectories of the long-term decline of coral reef ecosystems. *Science* **301**, 955–958 (2003).
37. Worm, B. & Tittensor, D. P. Range contraction in large pelagic predators. *Proc. Natl Acad. Sci. USA* **108**, 11942–11947 (2011).
38. McCauley, D. J. *et al.* Marine defaunation: animal loss in the global ocean. *Science* **347**, 1255641 (2015).  
**The proximate trigger for one of us (P.M.H.) to begin pondering the importance of rarity during events of geological proportion.**
39. Ceballos, G. & Ehrlich, P. R. Mammal population losses and the extinction crisis. *Science* **296**, 904–907 (2002).
40. Dornelas, M. *et al.* Assemblage time series reveal biodiversity change but not systematic loss. *Science* **344**, 296–299 (2014).
41. Hughes, J. B., Daily, G. C. & Ehrlich, P. R. Population diversity: its extent and extinction. *Science* **278**, 689–692 (1997).
42. Baum, J. K. *et al.* Collapse and conservation of shark populations in the Northwest Atlantic. *Science* **299**, 389–392 (2003).
43. Myers, R. A. & Worm, B. Rapid worldwide depletion of predatory fish communities. *Nature* **423**, 280–283 (2003).
44. Dulvy, N. K., Sadovy, Y. & Reynolds, J. D. Extinction vulnerability in marine populations. *Fish Fish.* **4**, 25–64 (2003).
45. Worm, B. *et al.* Impacts of biodiversity loss on ocean ecosystem services. *Science* **314**, 787–790 (2006).
46. Edgar, G. J. *et al.* Global conservation outcomes depend on marine protected areas with five key features. *Nature* **506**, 216–220 (2014).
47. Lotze, H. K., Coll, M., Magera, A. M., Ward-Paige, C. & Airoldi, L. Recovery of marine animal populations and ecosystems. *Trends Ecol. Evol.* **26**, 595–605 (2011).
48. Rabinowitz, D. in *The biological aspects of rare plant conservation* (ed. H. Synge) 205–217 (Wiley, 1981).
49. Sperling, E. A. in *Ecosystems Paleobiology and Geobiology. The Paleontological Society Papers* Vol. 19 (eds A. M. Bush, S. B. Pruss, & J. L. Payne) 77–86 (Geological Soc. America, 2013).
50. Schopf, T. J. M. Fossilization potential of an intertidal fauna: Friday Harbor, Washington. *Paleobiology* **4**, 261–270 (1978).
51. Briggs, D. E. G. The role of decay and mineralization in the preservation of soft-bodied fossils. *Annu. Rev. Earth Planet. Sci.* **31**, 275–301 (2003).
52. Benton, M. J. Biodiversity on land and in the sea. *Geol. J.* **36**, 211–230 (2001).
53. Benton, M. J. Diversification and extinction in the history of life. *Science* **268**, 52–58 (1995).
54. Nee, S. & May, R. M. Dynamics of metapopulations: habitat destruction and competitive coexistence. *J. Anim. Ecol.* **61**, 37–40 (1992).
55. Tilman, D. *et al.* Habitat destruction and the extinction debt. *Nature* **371**, 65–66 (1994).  
**The paper that defined extinction debt and made a strong case for the importance of events that occur long before the last individual dies in ecosystem change and extinction.**
56. Twitchett, R. J. Incompleteness of the Permian-Triassic fossil record: a consequence of productivity decline? *Geol. J.* **36**, 341–353 (2001).
57. Twitchett, R. J., Wignall, P. B. & Benton, M. J. Discussion on Lazarus taxa and fossil abundance at times of biotic crisis. *J. Geol. Soc. Lond.* **157**, 511–512 (2000).
58. Marshall, C. R. in *Quantitative Methods in Paleobiology* (eds Alroy, J. & Hunt, G.) 291–316 (The Paleontological Society, 2010).
59. Gardmark, A. *et al.* Regime shifts in exploited marine food webs: detecting mechanisms underlying alternative stable states using size-structured community dynamics theory. *Phil. Trans. R. Soc. Lond. B* **370**, 20130262 (2014).
60. deYoung, B. *et al.* Regime shifts in marine ecosystems: detection, prediction and management. *Trends Ecol. Evol.* **23**, 402–409 (2008).
61. Jackson, J. B. C. What was natural in the coastal oceans? *Proc. Natl Acad. Sci. USA* **98**, 5411–5418 (2001).
62. Rothschild, B. J., Ault, J. S., Gouletquer, P. & Heral, M. Decline of the Chesapeake Bay oyster population: a century of habitat destruction and overfishing. *Mar. Ecol. Prog. Ser.* **111**, 29–39 (1994).
63. Frank, K. T., Petrie, B., Choi, J. S. & Leggett, W. C. Trophic cascades in a formerly cod-dominated ecosystem. *Science* **308**, 1621–1623 (2005).
64. Jackson, J. B. C., Donovan, M. K., Cramer, K. L. & Lam, W. *Status and Trends of Caribbean Coral Reefs: 1970–2012*. (Global Coral Reef Monitoring Network, IUCN, 2014).
65. Levin, P. S. & Möllmann, C. Marine ecosystem regime shifts: challenges and opportunities for ecosystem-based management. *Phil. Trans. R. Soc. Lond. B* **370**, 20130275 (2014).
66. Wood, R. The changing biology of reef-building. *Palaos* **10**, 517–529 (1995).
67. Kiessling, W. & Simpson, C. On the potential for ocean acidification to be a general cause of ancient reef crises. *Glob. Change Biol.* **17**, 56–67 (2011).
68. Crane, P. R., Friis, E. M. & Pedersen, K. R. The origin and early diversification of angiosperms. *Nature* **374**, 27–33 (1995).
69. Edwards, E. J. *et al.*; C4 Grasses Consortium. The origins of C4 grasslands: integrating evolutionary and ecosystem science. *Science* **328**, 587–591 (2010).
70. Harries, P. J., Kauffman, E. G. & Hansen, T. A. in *Biotic Recovery from Mass Extinction Events. Geological Society of London Special Publication* 102 (ed. M. B. Hart) 41–60 (1996).
71. Kauffman, E. G. & Erwin, D. H. Surviving mass extinctions. *Geotimes* **40**, 14–17 (1995).
72. Jablonski, D. in *Dynamics of Extinction* (ed. Elliott, D. K.) 183–229 (Wiley, 1986).
73. Erwin, D. in *Evolutionary paleobiology* (eds Jablonski, D., Erwin, D. H. & Lipps, J. H.) 398–418 (Univ. Chicago Press, 1996).
74. Rong, J. Y., Boucot, A. J., Harper, D. A. T., Zhan, R. B. & Neuman, R. B. Global analyses of brachiopod faunas through the Ordovician and Silurian transition: reducing the role of the Lazarus effect. *Can. J. Earth Sci.* **43**, 23–39 (2006).
75. Surlyk, F. & Johansen, M. B. End-Cretaceous brachiopod extinctions in the chalk of Denmark. *Science* **223**, 1174–1177 (1984).
76. Casier, J. G. & Lethiers, F. Ostracods surviving the F/F event in the Devils Gate Pass Section (Nevada, USA). *Geobios* **30**, 811–821 (1997).
77. Smith, J. L. B. A living fish of Mesozoic type. *Nature* **143**, 455–456 (1939).
78. Hagino, K. *et al.* Re-discovery of a “living fossil” coccolithophore from the coastal waters of Japan and Croatia. *Mar. Micropaleontol.* **116**, 28–37 (2015).
79. Jablonski, D. Survival without recovery after mass extinctions. *Proc. Natl Acad. Sci. USA* **99**, 8139–8144 (2002).  
**The first detailed documentation of the importance of delayed extinctions across mass extinction boundaries.**
80. Jablonski, D. Lessons from the past: evolutionary impacts of mass extinctions. *Proc. Natl Acad. Sci. USA* **98**, 5393–5398 (2001).
81. Kaim, A. & Nutzel, A. Dead bellerophonitids walking - The short Mesozoic history of the Bellerophonitoidea (Gastropoda). *Palaogeogr. Palaeoclimatol. Palaeoecol.* **308**, 190–199 (2011).
82. Schubert, J. K. & Bottjer, D. J. Early Triassic stromatolites as post mass extinction disaster forms. *Geology* **20**, 883–886 (1992).
83. Ritterbush, K. A., Bottjer, D. J., Corsetti, F. A. & Rosas, S. New evidence on the role of siliceous sponges in ecology and sedimentary facies development in Eastern Panthalassa following the Triassic-Jurassic mass extinction. *Palaos* **29**, 652–668 (2014).
84. Pietsch, C. & Bottjer, D. J. The importance of oxygen for the disparate recovery patterns of the benthic macrofauna in the Early Triassic. *Earth Sci. Rev.* **137**, 65–84 (2014).
85. Peters, S. E. & Heim, N. A. in *Comparing the Geological and Fossil Records: Implications for Biodiversity Studies* (eds McGowan, A. J. & Smith, A. B.) 95–104 (Geological Society, 2011).
86. Smith, A. B., Lloyd, G. T. & McGowan, A. J. Phanerozoic marine diversity: rock record modelling provides an independent test of large-scale trends. *Proc. R. Soc. Lond. B* **279**, 4489–4495 (2012).
87. D'Hondt, S. Consequences of the Cretaceous/Paleogene mass extinction for marine ecosystems. *Annu. Rev. Ecol. Syst.* **36**, 295–317 (2005).
88. Hull, P. M. & Norris, R. D. Diverse patterns of ocean export productivity change across the Cretaceous-Paleogene boundary: New insights from biogenic barium. *Paleoceanography* **26**, 1–10 (2011).
89. Ward, P. D. *et al.* Sudden productivity collapse associated with the Triassic-Jurassic boundary mass extinction. *Science* **292**, 1148–1151 (2001).
90. Alegret, L., Thomas, E. & Lohmann, K. C. End-Cretaceous marine mass extinction not caused by productivity collapse. *Proc. Natl Acad. Sci. USA* **109**, 728–732 (2012).
91. Meyer, K. M., Yu, M., Jost, A. B., Kelley, B. M. & Payne, J. L.  $\delta^{13}\text{C}$  evidence that high primary productivity delayed recovery from end-Permian mass extinction. *Earth Planet. Sci. Lett.* **302**, 378–384 (2011).
92. Dayton, P. K., Tegner, M. J., Edwards, P. B. & Riser, K. L. Sliding baselines, ghosts, and reduced expectations in kelp forest communities. *Ecol. Appl.* **8**, 309–322 (1998).
93. Jackson, J. B. C. Ecological extinction and evolution in the brave new ocean. *Proc. Natl Acad. Sci. USA* **105** (Suppl. 1), 11458–11465 (2008).  
**A compelling case for ecological rarity in resetting ecosystems in the brave new oceans of the Anthropocene.**
94. Greenstein, B. J., Curran, H. A. & Pandolfi, J. M. Shifting ecological baselines and the demise of *Acropora cervicornis* in the western North Atlantic and Caribbean Province: a Pleistocene perspective. *Coral Reefs* **17**, 249–261 (1998).
95. Pandolfi, J. M. & Jackson, J. B. C. Ecological persistence interrupted in Caribbean coral reefs. *Ecol. Lett.* **9**, 818–826 (2006).  
**An elegant examination of resilience and collapse in coral reef communities, and an example of the potential of the fossil record to inform questions of conservation biology.**
96. Hanski, I. & Ovaskainen, O. Extinction debt at extinction threshold. *Conserv. Biol.* **16**, 666–673 (2002).

97. Smith, J. T. & Jackson, J. B. C. Ecology of extreme faunal turnover of tropical American scallops. *Paleobiology* **35**, 77–93 (2009).
98. Lewis, S. L. & Maslin, M. A. Defining the Anthropocene. *Nature* **519**, 171–180 (2015).
99. Crutzen, P. J. & Stoermer, E. F. The “Anthropocene”. *Global Change Newsletter IGBP* **41**, 17–18 (2000).
100. Zalasiewicz, J., Williams, M., Haywood, A. & Ellis, M. The Anthropocene: a new epoch of geological time? *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences* **369**, 835–841 (2011).
101. Steffen, W., Grinevald, J., Crutzen, P. & McNeill, J. The Anthropocene: conceptual and historical perspectives. *Philos. Trans. A* **369**, 842–867 (2011).
102. Schipper, J. *et al.* The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science* **322**, 225–230 (2008).
103. McKinney, M. L. High rates of extinction and threat in poorly studied taxa. *Conserv. Biol.* **13**, 1273–1281 (1999).
104. Régnier, C., Fontaine, B. & Bouchet, P. Not knowing, not recording, not listing: numerous unnoticed mollusk extinctions. *Conserv. Biol.* **23**, 1214–1221 (2009).
105. Erwin, D. A call to the custodians of deep time. *Nature* **462**, 282–283 (2009).
106. Peters, S. E. The Paleobiology Database Release PBDB Navigator. *Priscum* **21**, 1–2 (2014).
107. Finnegan, S. *et al.* Extinctions. Paleontological baselines for evaluating extinction risk in the modern oceans. *Science* **348**, 567–570 (2015).
108. Harnik, P. G., Simpson, C. & Payne, J. L. Long-term differences in extinction risk among the seven forms of rarity. *Proc. R. Soc. Lond. B* **279**, 4969–4976 (2012).
109. Benton, M. J. in *The unity of evolutionary biology* (ed. Dudley, E. C.) 89–102 (Dioscorides Press, 1991).
110. Barnosky, A. D. *et al.* Approaching a state shift in Earth's biosphere. *Nature* **486**, 52–58 (2012).
111. Lotze, H. K. & Worm, B. Historical baselines for large marine animals. *Trends Ecol. Evol.* **24**, 254–262 (2009).
112. Flessa, K. W. & Jablonski, D. Extinction is here to stay. *Paleobiology* **9**, 315–321 (1983).
113. Burgess, S. D., Bowring, S. & Shen, S. Z. High-precision timeline for Earth's most severe extinction. *Proc. Natl Acad. Sci. USA* **111**, 3316–3321 (2014).
114. Shen, S. Z. *et al.* Calibrating the end-Permian mass extinction. *Science* **334**, 1367–1372 (2011).
115. Schoene, B., Guex, J., Bartolini, A., Schaltegger, U. & Blackburn, T. J. Correlating the end-Triassic mass extinction and flood basalt volcanism at the 100 ka level. *Geology* **38**, 387–390 (2010).

**Acknowledgements** This manuscript arose out of discussion sparked by Arizona State University's Origins Project workshop hosted by L. Krauss and M. Laubichler; interdisciplinary training in the first class of the National Science Foundation IGERT programme in the Center for Marine Biodiversity & Conservation (led by N. Knowlton, J. B. C. Jackson, E. Sala, R. Carson, M. Tillman; supported by P. Dockery) at the Scripps Institution of Oceanography; and long association with D. E. G. Briggs and group. This manuscript was greatly improved through discussions with J. B. C. Jackson, K. L. Cramer, M. S. Roth and the Yale Paleontology group. D.H.E. acknowledges support from the NASA Astrobiology Institute. S.A.F.D. acknowledges support from a Peter Buck Fellowship at NMNH.

**Author Contributions** All authors contributed to the writing of this manuscript and the ideas contained therein.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.M.H. ([pincelli.hull@yale.edu](mailto:pincelli.hull@yale.edu)).



# Growth and splitting of neural sequences in songbird vocal development

Tatsuo S. Okubo<sup>1</sup>, Emily L. Mackevicius<sup>1</sup>, Hannah L. Payne<sup>2</sup>, Galen F. Lynch<sup>1</sup> & Michale S. Fee<sup>1</sup>

**Neural sequences are a fundamental feature of brain dynamics underlying diverse behaviours, but the mechanisms by which they develop during learning remain unknown. Songbirds learn vocalizations composed of syllables; in adult birds, each syllable is produced by a different sequence of action potential bursts in the premotor cortical area HVC. Here we carried out recordings of large populations of HVC neurons in singing juvenile birds throughout learning to examine the emergence of neural sequences. Early in vocal development, HVC neurons begin producing rhythmic bursts, temporally locked to a 'prototype' syllable. Different neurons are active at different latencies relative to syllable onset to form a continuous sequence. Through development, as new syllables emerge from the prototype syllable, initially highly overlapping burst sequences become increasingly distinct. We propose a mechanistic model in which multiple neural sequences can emerge from the growth and splitting of a common precursor sequence.**

Sequences of neural activity have been observed during various behaviours, including navigation<sup>1–4</sup>, short-term memory<sup>5–7</sup>, decision making<sup>8,9</sup>, and complex movements<sup>10,11</sup>, suggesting that neural sequences are a fundamental form of brain dynamics<sup>12,13</sup>. However, the circuit mechanisms underlying the generation of neural sequences and their development during learning are not well understood.

The songbird is a good model system to address such questions because the song produced by adults is learned during development<sup>14–18</sup>. Furthermore, adult song is associated with neural sequences in nucleus HVC<sup>19–24</sup>, a premotor cortical area necessary for the production of stereotyped adult song<sup>25–30</sup>. Most projection neurons in HVC generate a brief burst of spikes at one specific time in the song motif and different neurons are active at different times in the song<sup>19–24,30</sup>; thus, distinct syllable types are produced by largely non-overlapping neural sequences in HVC. Here we ask how these different neural sequences are constructed during vocal development.

Zebra finches acquire their stereotyped song through a gradual learning process<sup>14,31</sup>. Young birds initially produce a highly variable 'subsong'<sup>31</sup>, akin to human babbling<sup>15</sup>. Birds then enter the protosyllable stage as they begin to incorporate syllables of a characteristic ~100 ms duration<sup>32–35</sup>. This is followed by the gradual emergence of multiple syllable types<sup>32,33,36</sup>, and a final 'motif' stage in which syllables are produced in a reliable sequence. While HVC activity is not required for subsong<sup>27,34,35</sup>, it is required for song components in all later stages, including protosyllables, emerging syllable types, and adult song<sup>25–28,34,35</sup>.

## Developmental progression of HVC activity

To elucidate the mechanisms by which neural sequences in HVC develop, we recorded from populations of HVC projection neurons in juvenile and adult birds ( $n = 1,149$  neurons, 35 birds; Extended Data Fig. 1a). At all stages of vocal development, HVC projection neurons generated brief bursts of spikes during singing (Fig. 1a–c, Extended Data Fig. 1b, c). In the subsong stage ( $n = 12$  birds; defined by exponential distribution of syllable durations, before the emergence of protosyllables) roughly half the neurons generated bursts not temporally locked to syllable onsets (Extended Data Fig. 1d), while the other half produced bursts that tended to occur at a particular latency relative

to subsong syllable onsets (Fig. 1a and Extended Data Fig. 1e–i; 19/39 neurons exhibited syllable locking). The fraction of neurons locked to syllable onsets exhibited a gradual and significant increase throughout vocal development (Fig. 1f; correlation with song stage:  $r = 0.22$ ,  $P < 10^{-10}$ ; see Methods) until, in adult birds, virtually every projection neuron generated bursts precisely locked to syllables, as previously described<sup>19–24</sup>.

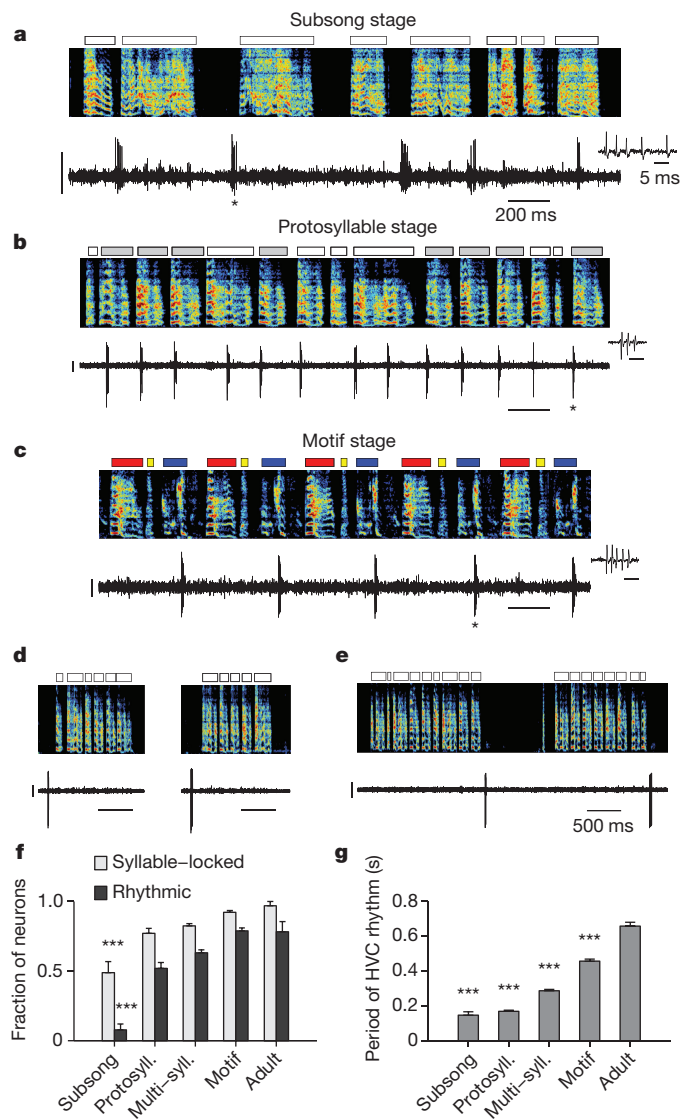
Song development is characterized by a gradual change in song rhythm<sup>33,37,38</sup>. The subsong stage, which has little evidence of rhythmic song structure, ends with the emergence of a rhythmically produced protosyllable (5–10 Hz)<sup>32–35</sup>. This is followed by a subsequent increase in the period between repetitions of the same sound, attributable to the addition of new song syllables<sup>33</sup>. HVC exhibited parallel changes in rhythmicity. In the subsong stage, most projection neurons did not burst rhythmically (Fig. 1a, f; 3/39 neurons were rhythmic). In the protosyllable stage, roughly half of the projection neurons generated rhythmic bursts (5–10 Hz) (Fig. 1b, f; 70/135 neurons were rhythmic; period  $169 \pm 6.4$  ms, mean  $\pm$  s.e.m.). Such bursts were typically locked to rhythmic protosyllables, but were also commonly observed during portions of the song with less rhythmic syllable onsets, particularly early in the protosyllable stage (Extended Data Fig. 2a–d). On average, both the fraction of rhythmic HVC neurons and the period of the HVC burst rhythm gradually increased during the emergence of new syllable types and the formation of the song motif (Fig. 1f, g; correlation between song stage and fraction of rhythmic neurons:  $r = 0.28$ ,  $P < 10^{-10}$ ; correlation between song stage and period of burst rhythm:  $r = 0.57$ ,  $P < 10^{-10}$ ).

A substantial fraction of projection neurons (285 of 1,117 neurons) in juvenile birds generated bursts related to song bouts—defined as epochs of continuous singing bounded by periods of silence (see Methods). Bout-related neurons generated brief bursts of spikes immediately before bout onset ('bout-onset' neurons; 137/285 neurons) or after bout offset (98/285 neurons) (Fig. 1d, e and Extended Data Fig. 2e–l; an additional 50/285 neurons were active both before and after bouts).

## Growth of a neural protosequence

We next wondered how the activity of HVC projection neurons is coordinated across the neural population during protosyllables. Multiple

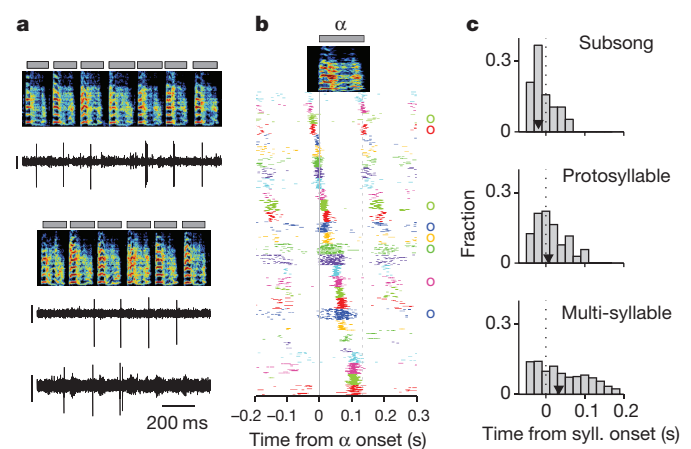
<sup>1</sup>McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. <sup>2</sup>Department of Neurobiology, Stanford University, Stanford, California 94305, USA.



**Figure 1 | Singing-related firing patterns of HVC projection neurons in juvenile birds.** **a**, Neuron recorded in the subsong stage, before the formation of protosyllables (RA-projecting HVC neuron, HVC<sub>RA</sub>; 51 dph; bird 7). Top, song spectrogram with syllables indicated above. Bottom, extracellular voltage trace. **b**, Neuron recorded in the protosyllable stage (HVC<sub>RA</sub>; 62 dph; bird 2). Protosyllables indicated (grey bars). **c**, Neuron recorded after motif formation (HVC<sub>RA</sub>; 68 dph; bird 8). **d**, Neuron bursting exclusively at bout onset (X-projecting HVC neuron, HVC<sub>X</sub>; 61 dph; bird 2). **e**, Neuron bursting exclusively at bout offset (HVC<sub>RA</sub>; 65 dph; bird 2). **f**, Developmental change in the fraction of neurons locked to syllable onsets (grey) and fraction of neurons with rhythmic bursting (black) (mean  $\pm$  s.e.m.;  $n = 39, 135, 565, 378$  and  $32$  neurons, respectively). **g**, Mean period of the HVC rhythmicity as a function of song stage ( $n = 3, 70, 356, 298$  and  $25$  neurons, respectively). \*\*\* $P < 0.001$ , post-hoc comparison with the adult stage. Spectrogram vertical axis  $500\text{--}8,000$  Hz. Scale bars for panels **a–c**,  $0.5$  mV,  $200$  ms; panels **d–e**,  $1$  mV,  $500$  ms. Inset in panels **a–c** show zoom of bursts indicated by an asterisk; scale bar,  $5$  ms.

recordings in the same bird revealed that different neurons were active at different times with respect to protosyllable onsets (Fig. 2a, b and Extended Data Figs 1n and 9k;  $n = 3$  birds,  $54$  neurons), with latencies spanning the duration of the protosyllable and the intervening gap ( $>90\%$  burst coverage; Extended Data Fig. 2t). These findings suggest that protosyllables are generated by a rhythmic protosequence—a repeating motor program comprised of a continuous sequence of bursts in HVC.

We next examined the developmental emergence of this rhythmic protosequence. In the subsong stage (Fig. 2c;  $n = 19$  neurons,  $12$  birds),



**Figure 2 | Rhythmic sequences in HVC during the protosyllable stage.** **a**, Three neurons recorded from bird 2 during protosyllable stage (top: HVC<sub>X</sub>; 63 dph; bottom: simultaneous recording two neurons; both HVC<sub>X</sub>; 64 dph; scale bar,  $0.5$  mV). **b**, Raster plot of  $28$  HVC projection neurons aligned to protosyllable onsets (sorted by latency;  $57\text{--}64$  dph, bird 2). Antidromically identified HVC<sub>RA</sub> neurons indicated by circles at right. **c**, Distribution of burst latencies relative to syllable onset in subsong stage (top), protosyllable stage (middle), and multi-syllable/motif stages (bottom), across all birds ( $n = 19, 104$  and  $814$  neurons, respectively). Black triangles indicate median burst times.

bursts had a significantly earlier distribution of latencies compared to the broader distribution of burst latencies in the protosyllable stage ( $n = 104$  neurons,  $13$  birds;  $P = 0.02$ ;  $63\%$  versus  $43\%$  of bursts before syllable onset in the subsong stage and protosyllable stage, respectively). Even though the range of latencies was narrower in subsong birds, different neurons recorded in the same bird were locked to syllable onsets at different latencies (Extended Data Fig. 1f–i). This suggests the existence of transient sequential activity, initiated just before syllable onset, but decaying within a few tens of milliseconds. This sequential activity appears to grow during the protosyllable stage to form longer sequences that can persist for more than a hundred milliseconds, throughout the duration of the protosyllable (Fig. 2b, c).

### Sequence splitting during syllable formation

We next wondered how distinct sequences in HVC, each corresponding to a distinct adult syllable type, emerge during vocal learning. Here we hypothesize that new syllable types can emerge by the gradual splitting of a single protosequence. In this view, we imagine that the neural sequences underlying newly emerging syllable types would initially be largely overlapping, with neurons shared across the emerging syllables. Splitting would be associated with an increasing number of neurons selective for a particular emerging syllable type, and a decreasing fraction of shared neurons.

To test this hypothesis, we recorded from HVC projection neurons ( $n = 769$ ) in  $6$  juvenile birds while they acquired multiple syllable types. As a first example, we will describe changes in the HVC population activity in a bird ( $n = 375$  projection neurons; bird 1) that developed two acoustically distinct syllable types (labelled  $\beta$  and  $\gamma$ ) over the course of several days (Fig. 3a, b;  $\beta$  and  $\gamma$  eventually form adult syllables B and C, respectively). During the protosyllable stage ( $56\text{--}59$  days post-hatch, dph), the majority of projection neurons participated in a rhythmic protosequence (Extended Data Fig. 1n;  $n = 14/16$  neurons; for example, Fig. 3c). After the emergence of syllable types  $\beta$  and  $\gamma$  ( $62\text{--}72$  dph), many neurons were selectively active only during  $\beta$  or during  $\gamma$ , but not both (Fig. 3d, f; of  $105$  neurons active during either  $\beta$  or  $\gamma$ ,  $41$  were  $\beta$ -specific and  $42$  were  $\gamma$ -specific). The bursts of these syllable-specific neurons exhibited a wide range of latencies, with spiking activity of neurons in each group spanning the entire duration of each syllable (Fig. 3g). Notably, we also observed a substantial population of neurons that were significantly active during both  $\beta$



and  $\gamma$  ( $n=22$  'shared' neurons; Fig. 3e–g). Simultaneous recordings revealed the co-occurrence, in different neurons, of shared and specific firing patterns (Fig. 3f, Extended Data Fig. 3a, b).

Shared neurons exhibited a number of striking characteristics. These neurons burst rhythmically with the same inter-burst interval as neurons recorded in the protosyllable stage (Fig. 3e, f; Extended Data Fig. 3f–j). Shared neurons were active, as a population, at a wide range of latencies within emerging syllables (Fig. 3g), and crucially, for a given shared neuron, the bursts during  $\beta$  occurred at a similar latency as the bursts during  $\gamma$  (Fig. 3g, Extended Data Fig. 4a–d). Thus, the population of shared neurons generated the same continuous burst sequence during both  $\beta$  and  $\gamma$ . This shared sequence occurred even at times when there was a significant acoustic difference between the shared syllables (Extended Data Fig. 5). We also found that the fraction of shared neurons later in development (81–112 dph) was significantly lower compared to the earlier recordings (Fig. 3h; 10 shared and 90 specific neurons;  $P=0.03$ ). Thus, the refinement of  $\beta$  and  $\gamma$  into the adult syllables B and C coincides with a decrease in the fraction of shared neurons, producing a gradual splitting of these representations into increasingly non-overlapping 'daughter' neural sequences.

The tendency of bird 1 to alternate between syllables  $\beta$  and  $\gamma$  means that syllable-specific neurons had an inter-burst interval, and thus a period, that was twice as long as that observed in the earlier protosyllable stage (Fig. 3c–f, Extended Data Fig. 3f–j). Therefore, the increase in the period of neural activity through skipping or alternating cycles of an underlying rhythm seems to be a basis for the increase in song period during vocal learning<sup>33</sup>.

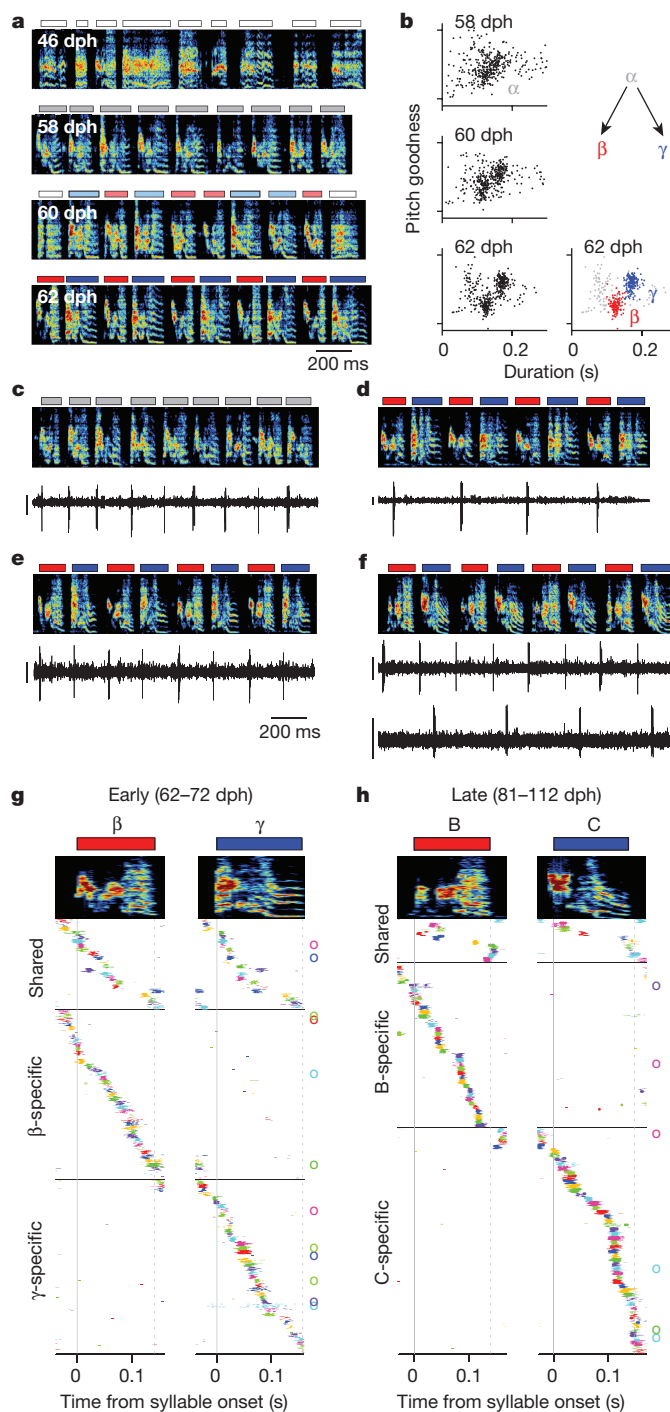
Although our key findings are described above for bird 1, a similar pattern of HVC coding by shared and specific neurons was seen in a total of 6 birds for which recordings were made during the emergence of multiple syllable types (birds 1–6; 185 shared neurons and 496 specific neurons for 8 syllable pairs analysed). Across three birds in which neurons were also recorded in later song stages, there was a significant decrease in the fraction of shared neurons during syllable development ( $n=5$  syllable pairs;  $P=3 \times 10^{-6}$ ; birds 1, 2 and 4). Neurons exhibiting an increased burst period by skipping cycles of an underlying rhythm were observed in 4 of the 6 birds (birds 1, 3, 4 and 6).

### Splitting in other learning strategies

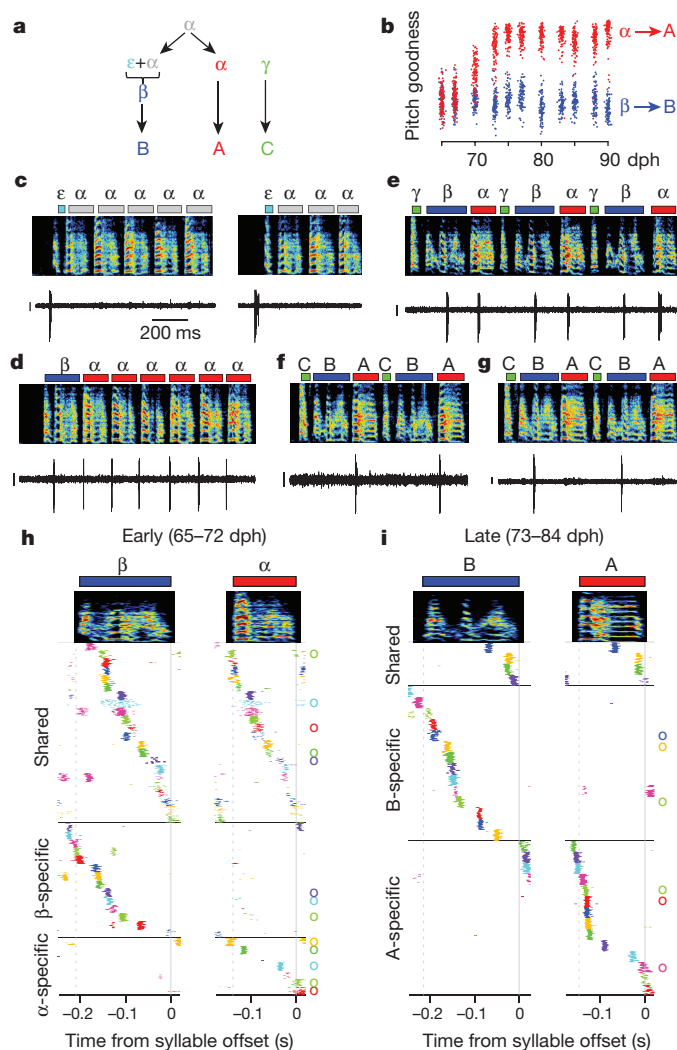
Behavioural studies have shown that new syllable types can emerge using several distinct developmental strategies<sup>32,33,36,39,40</sup>. The bird described above (bird 1) used the 'serial repetition' strategy<sup>32</sup> and 'sound differentiation *in situ*'<sup>33</sup> to develop two new syllables by alternating increasingly different variants of the protosyllable. Alternatively, birds can acquire multiple syllables simultaneously to form an entire motif ('motif strategy')<sup>32</sup>, or form new syllables at bout edges (onset or offset)<sup>39,40</sup>. We wondered if the splitting of neural sequences underlies these other strategies as well.

Neural recordings were obtained in three birds (birds 1, 2 and 5) that exhibited bout-onset syllable formation. We focus here on bird 2 in which projection neurons were recorded throughout song development (57–84 dph). Tracking of syllable structure (Extended Data Fig. 6) revealed that syllables A and B of the adult song derived from a common, rhythmically repeated protosyllable (labelled  $\alpha$ ; Fig. 4a, b), and that syllable B arose from the first repetition of  $\alpha$  at bout onset (Fig. 4c, d). The bout-onset syllable emerged as a distinct syllable type (labelled  $\beta$ ) by fusion of this first  $\alpha$  with a brief vocal element  $\epsilon$  at bout onset (Fig. 4c, d and Extended Data Fig. 6a–e).

To examine the neural mechanisms underlying the emergence of the new syllable  $\beta$  at bout onsets, we analysed the firing patterns of 125 HVC projection neurons. Before the emergence of syllable  $\beta$ , the majority of recorded projection neurons participated in a rhythmic protosequence (Fig. 2b;  $n=28/35$  neurons; 57–64 dph). A different subset of neurons was active at bout onsets (Fig. 4c; 4 of 35 neurons). After the reliable emergence of  $\beta$  at bout onsets, roughly half



**Figure 3 | Shared and specific sequences during the emergence of multiple syllable types.** All data are from bird 1. **a**, Song examples during the emergence of syllables  $\beta$  (red) and  $\gamma$  (blue). Panels show, from top to bottom, subsong stage (46 dph), rhythmic repetition of protosyllable  $\alpha$  (grey bars; 58 dph), rhythmic repetition of variants of the protosyllable ( $\beta$  and  $\gamma$ ; 60 dph), and further acoustic differentiation of  $\beta$  and  $\gamma$  (red and blue bars; 62 dph). **b**, Scatter plot of syllable duration versus mean pitch goodness (each dot is one syllable rendition;  $n=400$  syllables per day; unclassified syllables grey). **c**, Neuron recorded during protosyllable stage (HVC<sub>X</sub>; 56 dph). **d**,  $\beta$ -specific neuron (HVC<sub>X</sub>; 64 dph). **e**, Shared neuron active during both  $\beta$  and  $\gamma$  (HVC<sub>RA</sub>; 68 dph). **f**, Simultaneously recorded pair of HVC<sub>X</sub> neurons: shared neuron (top) and  $\gamma$ -specific neuron (bottom; 71 dph). **g**, Raster of 105 projection neurons early in syllable differentiation showing shared and specific sequences. HVC<sub>RA</sub> neurons indicated by circles at right. **h**, Same as **g** but for 100 neurons recorded after differentiation of  $\beta$  and  $\gamma$  into adult syllables B and C. Scale bars for panels c–f, 0.5 mV, all have the same time scale.



**Figure 4 | Shared and specific sequences during the emergence of a new syllable at bout onset.** All data are from bird 2. **a**, Schematic of syllable formation. **b**, Scatter plot of mean pitch goodness of syllables  $\alpha$  (red) and  $\beta$  (blue) through development ( $n = 100$  syllables per day; horizontal jitter added to improve data visibility). **c**, Bout-onset neuron active before element  $\epsilon$  (HVC<sub>RA</sub>; 64 dph). **d**, New syllable  $\beta$  formed by fusion of  $\epsilon$  and  $\alpha$ . Neuron shared between  $\alpha$  and  $\beta$  (HVC<sub>RA</sub>; 65 dph). **e**, Neuron shared between  $\alpha$  and  $\beta$  (HVC<sub>X</sub>; 70 dph). **f**, A-specific neuron (HVC<sub>RA</sub>; 80 dph). **g**, B-specific neuron (HVC<sub>RA</sub>; 73 dph). **h**, Population raster plot of 43 projection neurons recorded early in the emergence of syllable  $\beta$  showing shared and specific sequences. **i**, Raster plot of 32 neurons recorded after differentiation of  $\beta$  and  $\alpha$  into adult syllables B and A. Scale bars for panels c–g, 0.5 mV, all have the same time scale.

of projection neurons generated bursts during both syllables  $\alpha$  and  $\beta$  (65–72 dph; Fig. 4d, e;  $n = 22$  ‘shared’ neurons; 21 ‘specific’ neurons). These shared neurons produced nearly identical sequences during these two syllables (Fig. 4h, Extended Data Fig. 4c). Later in song development (73–84 dph), we observed a smaller fraction of shared neurons ( $n = 4$  ‘shared’ neurons;  $P = 5 \times 10^{-4}$ ), and a correspondingly larger fraction of syllable-specific neurons (Fig. 4f, g, i;  $n = 28$  ‘specific’ neurons), consistent with a gradual splitting of the proto-sequence into increasingly non-overlapping ‘daughter’ sequences. Evidence for sequence splitting during bout-onset differentiation was also observed in birds 1 and 5 (Extended Data Fig. 7).

Note that the bout-onset differentiation in bird 1 occurred after the earlier emergence of the syllables  $\beta$  and  $\gamma$  (Fig. 3), suggesting that new syllables may emerge in a hierarchical process—that is, by the splitting of sequences that are themselves the product of an earlier splitting process (Extended Data Fig. 7).

We were able to examine the question of whether neural sequence splitting also underlies the ‘motif strategy’ of song learning in two birds (birds 3 and 4; Extended Data Figs 8 and 9). In both birds, neural recordings showed the existence of rhythmically bursting neurons in the protosyllable stage (Extended Data Figs 8e and 9e, f). After the emergence of multiple syllable types, every syllable in the emerging motif had at least one neuron that was shared with another syllable at similar latencies (Extended Data Figs 8f–j and 9g–o), consistent with the view that all of these syllables arose from the simultaneous splitting of a common protosequence.

## Mechanistic model and discussion

We propose a mechanistic model of learning in the HVC network to describe how sequences emerge during song development. This model is based on the idea that sequential bursting results from the propagation of activity through a continuous synaptically connected chain of neurons within HVC<sup>21,41–47</sup>. It also captures non-uniformities such as increased burst density at syllable onsets, as formulated in a perspective of HVC function emphasizing vocal gestures<sup>22</sup>.

Modelling studies have shown that a combination of two synaptic plasticity rules—spike-timing dependent plasticity (STDP) and heterosynaptic competition—can transform a randomly connected network into a feedforward synaptically connected chain that generates sparse sequential activity<sup>43,44</sup>. We hypothesize that the same mechanisms can drive the formation of a rhythmic protosyllable chain, and subsequently split this chain into multiple daughter chains for different syllable types. To test this hypothesis, we constructed a simple network of binary units representing HVC projection neurons<sup>44</sup>.

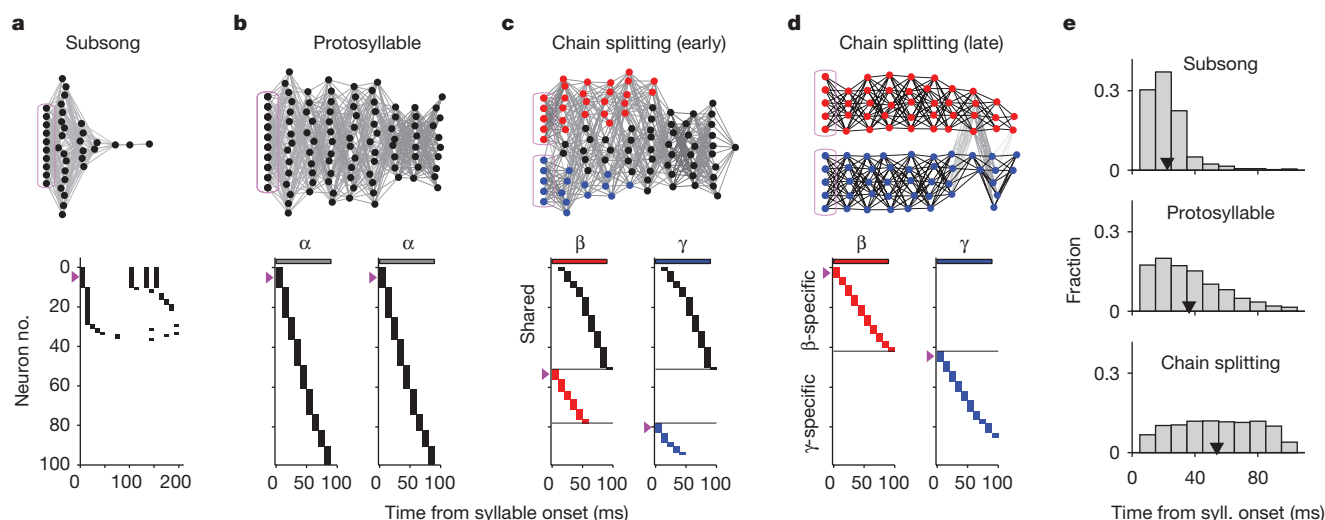
The model neurons are initially connected with random excitatory weights, representing the subsong stage. We hypothesize that a subset of HVC neurons receives an external input at syllable onsets and serves as a seed from which chains grow during later learning stages<sup>43,45</sup>. Before learning, activation of these seed neurons produced a transiently propagating sequence of network activity that decayed rapidly (within tens of milliseconds; Fig. 5a).

In the next stage, the network is trained to produce a single protosyllable by activating seed neurons rhythmically (100 ms period). The connections are modified according to the learning rules described above<sup>43,44</sup>. As a result, connections were strengthened along the population of neurons sequentially activated after syllable onsets, resulting in the growth of a feedforward synaptically connected chain that supported stable propagation of activity (Fig. 5b).

We found that this single chain could be induced to split into two daughter chains by dividing the seed neurons into two groups that were activated on alternate cycles of the rhythm (Fig. 5c, d and Supplementary Video 1). Local inhibition<sup>48</sup> and synaptic competition were also increased (see Methods). During the splitting process, we observed neurons specific to each of the emerging syllable types, as well as shared neurons that were active at the same latencies in both syllable types (Fig. 5c). Just as observed in our data, over the course of development the distribution of burst latencies in the model continued to broaden (Fig. 5e), and the fraction of shared neurons decreased (Fig. 5c, d). The average period of rhythmic bursting in model neurons increased during chain splitting as neurons became ‘specific’ for one emerging syllable type and began to participate only on alternate cycles of the protosyllable rhythm (Fig. 5d and Extended Data Fig. 10g, h).

Our model can reproduce other strategies by which birds learn new syllable types. We implemented bout-onset differentiation in the model by also including a population of seed neurons activated at bout onsets (see Figs 1d and 4c, and Extended Data Fig. 10a). This caused the protosyllable chain to split in such a way that one daughter chain was reliably activated only at bout onsets, while the other daughter chain was active only on subsequent syllables (Extended Data Fig. 10a–d and Supplementary Video 2). Our model was also able to simulate the simultaneous emergence of a three-syllable motif (‘motif





**Figure 5 | A neural model of sequence formation and splitting in HVC.** **a–d**, Top, network diagrams of participating neurons (darker lines indicate stronger connections; magenta boxes indicate seed neurons). Bottom, raster plot of neurons showing shared and specific sequences. Neurons sorted by relative latency. Magenta arrows indicate groups of seed neurons. **a**, Subsong stage: activation of seed neurons produces a rapidly decaying burst of sequential activity. **b**, Protosyllable stage: rhythmic activation of

seed neurons induces formation of a protosyllable chain. **c**, Alternating activation of red and blue seed neurons and synaptic competition drives the network to split into two chains (specific neurons, red and blue; shared neurons, black). **d**, Network after chain splitting. **e**, Distribution of model burst latencies during subsong, protosyllable stage and chain splitting stage (early and late combined).

strategy’) by dividing the seed neurons into three subpopulations (Extended Data Fig. 10e–h).

Our data and modelling support the possibility of syllable formation by mechanisms other than sequence splitting. For example, in several birds, a short vocal element emerged at bout onsets that did not seem to differentiate acoustically from the protosyllable (and thus was not bout-onset differentiation; for example, ‘E’ in bird 1, Extended Data Fig. 7a; or ‘C’ in bird 2, Extended Data Fig. 6a, b). We found that, by using different learning parameters, our model allows bout-onset seed neurons to induce the formation of a new syllable chain at bout onset, rather than inducing bout-onset differentiation (Extended Data Fig. 10i–k).

In summary, our model of learning in a simple sequence-generating network captures transformations that underlie the formation of new syllable types via a diverse set of learning strategies.

### Possible role of sequence splitting

The process of splitting a prototype neural sequence allows learned components of a prototype motor program to be reused in each of the daughter motor programs. For example, one of the earliest aspects of vocal learning is the coordination between singing and breathing<sup>35</sup>, specifically, the alternation between vocalized expiration and non-vocalized inspiration typical of adult song<sup>49</sup>. The protosequence in HVC would allow the bird to learn the appropriate coordination of respiratory and vocal musculature. Duplication of the protosequence through splitting would result in two ‘functional’ daughter sequences, each already capable of proper vocal/respiratory coordination, and each suitable as a substrate for rapid learning of a new syllable type.

This proposed mechanism resembles a process thought to underlie the evolution of novel gene functions: gene duplication followed by divergence through independent mutations<sup>50</sup>. Similarly, for the acquisition of complex behaviours, the duplication of neural sequences by splitting, followed by independent differentiation through learning, may provide a mechanism for constructing complex motor programs.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 January; accepted 22 September 2015.

Published online 30 November 2015.

- Wikenheiser, A. M. & Redish, A. D. Hippocampal theta sequences reflect current goals. *Nature Neurosci.* **18**, 289–294 (2015).
- Pfeiffer, B. E. & Foster, D. J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).
- Dragoi, G. & Tonegawa, S. Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature* **469**, 397–401 (2011).
- Davidson, T. J., Kloosterman, F. & Wilson, M. A. Hippocampal replay of extended experience. *Neuron* **63**, 497–507 (2009).
- Fujisawa, S., Amarasingham, A., Harrison, M. T. & Buzsáki, G. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature Neurosci.* **11**, 823–833 (2008).
- Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsáki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321**, 1322–1327 (2008).
- Eichenbaum, H. Time cells in the hippocampus: a new dimension for mapping memories. *Nature Rev. Neurosci.* **15**, 732–744 (2014).
- Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484**, 62–68 (2012).
- Murakami, M., Vicente, M. I., Costa, G. M. & Mainen, Z. F. Neural antecedents of self-initiated actions in secondary motor cortex. *Nature Neurosci.* **17**, 1574–1582 (2014).
- Peters, A. J., Chen, S. X. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510**, 263–267 (2014).
- Tanji, J. Sequential organization of multiple movements: involvement of cortical motor areas. *Annu. Rev. Neurosci.* **24**, 631–651 (2001).
- Buzsáki, G. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* **68**, 362–385 (2010).
- Vogels, T. P., Rajan, K. & Abbott, L. F. Neural network dynamics. *Annu. Rev. Neurosci.* **28**, 357–376 (2005).
- Immelmann, K. in *Bird Vocalizations* (ed. Hinde, R. A.) 61–74 (Cambridge Univ. Press, 1969).
- Doupe, A. J. & Kuhl, P. K. Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* **22**, 567–631 (1999).
- Mooney, R. Neural mechanisms for learned birdsong. *Learn. Mem.* **16**, 655–669 (2009).
- Konishi, M. Birdsong: from behavior to neuron. *Annu. Rev. Neurosci.* **8**, 125–170 (1985).
- Brainard, M. S. & Doupe, A. J. Translating birdsong: songbirds as a model for basic and applied medical research. *Annu. Rev. Neurosci.* **36**, 489–517 (2013).
- Hahnloser, R. H., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419**, 65–70 (2002).
- Kozhevnikov, A. A. & Fee, M. S. Singing-related activity of identified HVC neurons in the zebra finch. *J. Neurophysiol.* **97**, 4271–4283 (2007).
- Long, M. A., Jin, D. Z. & Fee, M. S. Support for a synaptic chain model of neuronal sequence generation. *Nature* **468**, 394–399 (2010).
- Amador, A., Perl, Y. S., Mindlin, G. B. & Margoliash, D. Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature* **495**, 59–64 (2013).
- Fujimoto, H., Hasegawa, T. & Watanabe, D. Neural coding of syntactic structure in learned vocalizations in the songbird. *J. Neurosci.* **31**, 10023–10033 (2011).
- Prather, J. F., Peters, S., Nowicki, S. & Mooney, R. Precise auditory-vocal mirroring in neurons for learned vocal communication. *Nature* **451**, 305–310 (2008).

25. Nottebohm, F., Stokes, T. M. & Leonard, C. M. Central control of song in the canary, *Serinus canarius*. *J. Comp. Neurol.* **165**, 457–486 (1976).
26. Long, M. A. & Fee, M. S. Using temperature to analyse temporal dynamics in the songbird motor pathway. *Nature* **456**, 189–194 (2008).
27. Aronov, D., Andalman, A. S. & Fee, M. S. A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science* **320**, 630–634 (2008).
28. Simpson, H. B. & Vicario, D. S. Brain pathways for learned and unlearned vocalizations differ in zebra finches. *J. Neurosci.* **10**, 1541–1556 (1990).
29. Ali, F. *et al.* The basal ganglia is necessary for learning spectral, but not temporal, features of birdsong. *Neuron* **80**, 494–506 (2013).
30. Vallentin, D. & Long, M. A. Motor origin of precise synaptic inputs onto forebrain neurons driving a skilled behavior. *J. Neurosci.* **35**, 299–307 (2015).
31. Zann, R. A. *The Zebra Finch: A Synthesis of Field and Laboratory Studies* (Oxford Univ. Press, 1996).
32. Liu, W. C., Gardner, T. J. & Nottebohm, F. Juvenile zebra finches can use multiple strategies to learn the same song. *Proc. Natl Acad. Sci. USA* **101**, 18177–18182 (2004).
33. Tchernichovski, O., Mitra, P. P., Lints, T. & Nottebohm, F. Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science* **291**, 2564–2569 (2001).
34. Aronov, D., Veit, L., Goldberg, J. H. & Fee, M. S. Two distinct modes of forebrain circuit dynamics underlie temporal patterning in the vocalizations of young songbirds. *J. Neurosci.* **31**, 16353–16368 (2011).
35. Veit, L., Aronov, D. & Fee, M. S. Learning to breathe and sing: development of respiratory-vocal coordination in young songbirds. *J. Neurophysiol.* **106**, 1747–1765 (2011).
36. Tchernichovski, O. & Mitra, P. P. Towards quantification of vocal imitation in the zebra finch. *J. Comp. Physiol. A* **188**, 867–878 (2002).
37. Glaze, C. M. & Troyer, T. W. Development of temporal structure in zebra finch song. *J. Neurophysiol.* **109**, 1025–1035 (2013).
38. Saar, S. & Mitra, P. P. A technique for characterizing the development of rhythms in bird song. *PLoS One* **3**, e1461 (2008).
39. Lipkind, D. *et al.* Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature* **498**, 104–108 (2013).
40. Lipkind, D. & Tchernichovski, O. Quantification of developmental birdsong learning from the subsyllabic scale to cultural evolution. *Proc. Natl Acad. Sci. USA* **108** (Suppl. 3), 15572–15579 (2011).
41. Jin, D. Z., Ramazanoglu, F. M. & Seung, H. S. Intrinsic bursting enhances the robustness of a neural network model of sequence generation by avian brain area HVC. *J. Comput. Neurosci.* **23**, 283–299 (2007).
42. Li, M. & Greenside, H. Stable propagation of a burst through a one-dimensional homogeneous excitatory chain model of songbird nucleus HVC. *Phys. Rev. E* **74**, 011918 (2006).
43. Jun, J. K. & Jin, D. Z. Development of neural circuitry for precise temporal sequences through spontaneous activity, axon remodeling, and synaptic plasticity. *PLoS One* **2**, e723 (2007).
44. Fiete, I. R., Senn, W., Wang, C. Z. & Hahnloser, R. H. Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity. *Neuron* **65**, 563–576 (2010).
45. Buonomano, D. V. A learning rule for the emergence of stable dynamics and timing in recurrent networks. *J. Neurophysiol.* **94**, 2275–2283 (2005).
46. Gibb, L., Gentner, T. Q. & Abarbanel, H. D. Inhibition and recurrent excitation in a computational model of sparse bursting in song nucleus HVC. *J. Neurophysiol.* **102**, 1748–1762 (2009).
47. Bertram, R., Daou, A., Hyson, R. L., Johnson, F. & Wu, W. Two neural streams, one voice: pathways for theme and variation in the songbird brain. *Neuroscience* **277**, 806–817 (2014).
48. Kosche, G., Vallentin, D. & Long, M. A. Interplay of inhibition and excitation shapes a premotor neural sequence. *J. Neurosci.* **35**, 1217–1227 (2015).
49. Goller, F. & Cooper, B. G. Peripheral motor dynamics of song production in the zebra finch. *Ann. NY Acad. Sci.* **1016**, 130–152 (2004).
50. Ohno, S. *Evolution by Gene Duplication* (Springer-Verlag, 1970).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank M. Wilson, J. Kornfeld, M. Jazayeri, S. Seung, N. Ji, and M. Stetner for comments on the manuscript. Funding to M.S.F. was provided by the NIH (grant no. R01DC009183) and by the Mathers Foundation, to T.S.O. by the Nakajima Foundation and Schoemaker Fellowship, to E.L.M. by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program, and to H.L.P. by the National Science Foundation (NSF) Graduate Research Fellowship Program (no. DGE-114747) and the NSF Integrative Graduate Education and Research Traineeship (no. 0801700). The modelling work was begun in the Methods in Computational Neuroscience course at the Marine Biological Laboratory (NIH grant number R25MH062204).

**Author Contributions** The study was conceived and designed by T.S.O. and M.S.F. Experimental data were collected by T.S.O. Data were analysed by T.S.O. and M.S.F. with contributions from G.F.L. The modelling study was performed by E.L.M. and H.L.P. in collaboration with T.S.O. and M.S.F. All five authors contributed to writing the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.S.F. ([fee@mit.edu](mailto:fee@mit.edu)).



## METHODS

**Animals.** We used juvenile male zebra finches (*Taeniopygia guttata*) 44–112 days post-hatch (dph) singing undirected song ( $n = 32$  birds). Animals were not divided into experimental groups; thus, randomization and blinding were not necessary. No statistical methods were used to predetermine sample size. Birds were obtained from the Massachusetts Institute of Technology zebra finch breeding facility (Cambridge, Massachusetts). The care and experimental manipulation of the animals were carried out in accordance with guidelines of the National Institutes of Health and were reviewed and approved by the Massachusetts Institute of Technology Committee on Animal Care.

All the juvenile birds were raised by their parents in individual breeding cages until  $38 \pm 5.2$  dph (mean  $\pm$  s.d.) when they were removed and were singly housed in custom-made sound isolation chambers (maintained on a 12:12 h day-night schedule). For a subset of the birds (birds 1, 2 and 4), additional tutoring was carried out after removal from the breeding cages to facilitate song imitation. This was done by playback of the tutor song through a speaker (20 bouts per day). Additional tutoring was done for 12 days for bird 1, 7 days for bird 2, and 18 days for bird 4. Bird identification key: bird 1, to3965; bird 2, to3779; bird 3, to3017; bird 4, to5640; bird 5, to3396; bird 6, to2309; bird 7, to3412; bird 8, to3567; bird 9, to2462; bird 10, to2331; bird 11, to2427; bird 12, to3352.

To compare the activity of HVC projection neurons in juvenile birds with that of adult birds, we also included neurons recorded in adults ( $>120$  dph,  $n = 3$  birds) which included a reanalysis of previously published HVC recordings performed in adult male zebra finches singing directed song<sup>20</sup>.

**Song recordings.** Songs were recorded with Sound Analysis Pro<sup>51</sup> or a custom-written MATLAB software (A. Andalman), which was configured to ensure triggering of recordings on all quiet vocalizations of juvenile birds<sup>27</sup>. The vertical axis range for all spectrograms is 500–8,000 Hz.

**Classification of song stages.** We classified each day of juvenile singing into one of four song stages: subsong stage, protosyllable stage, multi-syllable stage, and motif stage (Extended Data Fig. 1a). Subsong stage ( $48 \pm 4$  dph, median  $\pm$  inter-quartile range, IQR) is defined as having a syllable duration distribution well-fit by an exponential distribution<sup>34,35</sup>, with an upper limit for the Lilliefors goodness-of-fit statistic of 6. Following the subsong stage, birds enter the protosyllable stage ( $58 \pm 10$  dph, median  $\pm$  IQR) characterized by the presence of syllables with consistent timing reflected in a peak in the distribution of syllable durations<sup>32–35</sup>. The onset of the protosyllable stage was defined here as the first day in which the syllable duration distribution deviated from an exponential distribution (Lilliefors goodness-of-fit statistic greater than 6). Following the protosyllable stage, birds transition to the multi-syllable stage ( $62 \pm 12$  dph, median  $\pm$  IQR) in which multiple distinct syllable types are visible in the song spectrogram and as multiple clusters in a scatter plot of syllable features<sup>52</sup> (for example, Fig. 3a, b; 62 dph). The motif stage ( $73 \pm 21$  dph, median  $\pm$  IQR) was defined by the production of a sequence of syllables in a relatively fixed order<sup>31</sup>. Finally, songs recorded in birds older than 120 dph were assigned as adult stage. A slightly older cutoff than the typical definition of adulthood in zebra finches ( $\sim 90$  dph)<sup>14</sup> was used, because some of our birds in the 90–120 dph range continued to undergo some small developmental changes, as has been reported<sup>21</sup>.

**Syllable segmentation and bout extraction.** Syllable segmentation of the juvenile song was done based on the song power in a spectral band between 1 and 4 kHz, as described previously<sup>27,34,35</sup>. In a few cases, cutoff frequencies of the band-pass filters were adjusted to avoid the inclusion of high-frequency inspiratory sounds<sup>35,53</sup>. Introductory notes were removed manually to avoid including HVC neurons that are rhythmically active during these elements<sup>54</sup>. Song bouts were defined as continuous sequences of syllables separated by gaps no longer than 300 ms<sup>35</sup>. Bout onset was defined as the onset of the first syllable in the bout, and bout offset was defined as the offset of the last syllable in the bout.

**Syllable segmentation based on the song rhythmicity ('phase segmentation').** For bird 3 ('motif strategy'), it was difficult to segment syllables consistently using previous methods based on setting a threshold on the sound amplitude<sup>27,34,35</sup>. To overcome this limitation, we segmented syllables based on the phase of the rhythmicity in the song ('phase segmentation'). The peak of the song rhythm, defined as the spectrum of the sound amplitude during singing<sup>38</sup>, exhibited a peak around 9 Hz (Extended Data Fig. 8c). To estimate the instantaneous phase of this rhythm, we first band-pass filtered the sound amplitude (Extended Data Fig. 8c, d; second-order IIR resonator filter with peak at 9 Hz and  $-3$  dB half-bandwidth of 3 Hz; MATLAB command `iirpeak`). The band-pass filtered signal was then processed using the Hilbert transform (MATLAB command `hilbert`) to compute the instantaneous amplitude and phase (Extended Data Fig. 8d). Next, we set a threshold on this instantaneous amplitude to find the rhythmic part of the song. Finally, within this rhythmic part, song was segmented by detecting threshold crossings of the instantaneous phase (Extended Data Fig. 8d, bottom). Phase

segments that contain no sounds or calls were manually removed. Similarly, phase segmentation (band-pass filter with peak at 10 Hz and half-bandwidth of 3 Hz) was used to segment the song during the protosyllable stage for bird 4 (Extended Data Fig. 9a, e, f). Note that this method is best suited for segmenting songs that have strong rhythmic modulation of song amplitude, but in which syllable boundaries are not strongly rhythmic. This appeared to be typical of birds employing the 'motif strategy'<sup>32</sup>.

**Syllable classification and labelling.** Protosyllables were defined by their characteristic durations as has been described previously<sup>34,35</sup>. In short, to identify the protosyllables, we first subtracted the best-fit exponential distribution (using 200–400 ms) from the syllable duration distribution, and fitted a Gaussian distribution to this residual. Protosyllables were defined as syllables having durations within two standard deviations from the mean of this Gaussian distribution. We labelled protosyllables using the Greek letter ' $\alpha$ ' in all our birds for consistency.

To label the emerging syllables in the juvenile song, we used the Greek letters  $\beta$ ,  $\gamma$ ,  $\delta$ , and  $\epsilon$ . In contrast, to label the syllables in the adult motif, we used the capital letters of the Latin alphabet A, B, C, etc. For birds in which the song learning trajectory was tracked developmentally, we labelled the syllables such that the correspondence between the juvenile syllables and adult syllables is straightforward: for example,  $\alpha$  becomes A,  $\beta$  becomes B,  $\gamma$  becomes C,  $\delta$  becomes D, and  $\epsilon$  becomes E. Note that this labelling scheme leads to a slightly unconventional labelling of adult song in the sense that a motif can have letters in a reverse order (for example, CBA in Fig. 4f, g; Extended Data Fig. 6a), or a motif might not have a syllable A (for example, EDCB in Extended Data Fig. 7a).

Syllable labelling was done manually by visual inspection of the song spectrogram; this was done blind with respect to the neural activity. The existence of multiple distinct syllable types were confirmed by calculating the syllable duration and acoustic features commonly used to analyse birdsong syllables<sup>51,55</sup>, and visualizing the clusters of syllables in a two-dimensional space<sup>52</sup> (Fig. 3b, Extended Data Figs 8b and 9d). In some cases, syllable order was used as an additional indicator of syllable identity (for example, Extended Data Fig. 7a, 70 dph; Extended Data Fig. 8a, 51 dph; Extended Data Fig. 9a, 59 dph).

In bird 1, syllables  $\beta$  and  $\gamma$  were labelled manually by visual inspection of the song spectrogram (Fig. 3a). Since characterizing shared neurons and specific neurons depends on the reliable labelling of syllables, we took a conservative approach and only labelled syllables that were clearly identifiable and did not label the syllables that were ambiguous (fraction of syllables labelled as  $\beta$  or  $\gamma$  during 62–66 dph:  $70 \pm 5.5\%$ , mean  $\pm$  s.d.). We then estimated the error rate of our labelling procedure by plotting the labelled syllables ( $n = 200$  syllables per type on each day) in a two-dimensional space of syllable duration and mean pitch goodness (Fig. 3b), and obtained a decision boundary using linear discriminant analysis. We used mismatch between manual labelling and feature-based labelling to estimate the error rate for syllables  $\beta$  and  $\gamma$ . The error rate during the first five days of syllable differentiation (62–66 dph), when the labelling was most difficult, was only 1.1% on average (range: 0.25–3.0%).

For the second round of differentiation in bird 1, syllable order was used to assist in the labelling of syllables in early stages when syllables 'B' and 'D' were not easily distinguishable based on acoustic differences. Because these syllables underwent bout-onset differentiation, the first  $\beta$  after bout onset was labelled 'D'; later renditions of  $\beta$  in the bout were labelled 'B' (Extended Data Fig. 7a).

In bird 2, several emerging syllables could be easily distinguished based on syllable durations (Extended Data Fig. 6d). Specifically, syllables whose durations were 110–160 ms, and 180–250 ms were defined as  $\alpha$  and  $\beta$ , respectively. Syllables that were 10–75 ms in duration were labelled  $\gamma$  if they were followed by a  $\beta$ , and labelled  $\epsilon$  otherwise.

**Chronic neural recordings.** Single-unit recordings of HVC projection neurons during singing were carried out using a motorized microdrive described previously<sup>56,57</sup>. Single-units were confirmed by the existence of the refractory period in the inter-spike interval (ISI) distribution (Extended Data Fig. 1b). Neurons that were active only during distance calls and not during singing<sup>20</sup> were excluded from the analysis. In addition, neurons recorded for less than 5 s of singing were excluded since the short recording duration did not allow us to reliably quantify the activity pattern of these neurons.

Antidromic identification of HVC projection neurons was carried out with a bipolar stimulating electrode implanted in RA and Area X (single pulse of 200  $\mu$ s every 1 s; current amplitude: 50–500  $\mu$ A)<sup>19,20,57–59</sup>. A subset of antidromically identified projection neurons was further validated with collision testing<sup>19,20,57–59</sup>. A different subset of single units were identified as putative projection neurons based on sparse bursting, but could not be antidromically identified because they did not respond to antidromic stimulation or were lost before antidromic identification could be carried out (211 of 1,149 neurons). These neurons were included in the data set as unidentified HVC projection neurons (HVC<sub>p</sub>).

**Analysis of neural activity.** Spikes were sorted offline using custom MATLAB software (D. Aronov).

**Definition of bursts.** HVC projection neurons exhibited bursts of action potentials during singing (Fig. 1a–c). The bursting nature of these neurons was evident in the inter-spike interval (ISI) distribution during singing, which exhibited two peaks with an inter-peak minimum near 30 ms (Extended Data Fig. 1b). We defined a ‘burst’ as a continuous group of spikes separated by intervals of 30 ms or less. Thus, by definition, bursts are separated from other spikes by intervals greater than 30 ms. Note that single spikes separated by more than 30 ms from both the preceding spike and the following spikes were also counted as a burst. Burst time was defined as the centre of mass of all the spikes within the burst. Burst width was defined as the interval between the first and the last spike in a burst (Extended Data Fig. 1c, top). Firing rate during burst was defined as the reciprocal of the mean inter-spike interval in a burst (Extended Data Fig. 1c, bottom). For the calculation of burst width and firing rate during bursts, bursts composed of a single spike were excluded.

**Syllable-related neural activity.** To analyse the temporal relation between neural activity and song syllables, we aligned the spike times to syllable onsets and constructed a rate histogram (1 ms bin, smoothed over 20 bins; range:  $\pm 0.5$  s from syllable onsets). The peak in this rate histogram was found between 50 ms before syllable onset and 200 ms after syllable onset. To test the significance of this peak, surrogate histograms were created by adding different random time shifts to the spike times on each trial<sup>60</sup>. Random time shifts were drawn from a uniform distribution over  $\pm 0.5$  s. The peak of this surrogate histogram was recorded, and this shuffling procedure was repeated 1,000 times; *P* values were obtained by analysing the frequency with which the peaks of surrogate data were larger than that of the real data, and  $P < 0.05$  was considered significant.

To visualize the population activity associated with protosyllables, we constructed a population raster plot by choosing 20 protosyllable renditions for which each neuron was most active. Different neurons were plotted in different colours (Fig. 2b, Extended Data Figs 1n and 9k). For all the other population raster plots associated with identified syllables, 20 random renditions were chosen for display. For all population raster plots, syllable duration from each rendition was linearly time-warped to the mean duration of the syllable. Spike times were warped by the same factor.

**Bout-related neural activity.** A subset of HVC projection neurons exhibited bout-related activity: bursting before bout onsets and/or after bout offsets (Fig. 1d, e and Extended Data Fig. 2e–l). To quantify the pre-bout activity, we generated histograms aligned to bout onsets (Extended Data Fig. 2f, g) and found the peak in the histogram in a 300 ms window before bout onset. We considered a neuron to be exhibiting ‘pre-bout activity’ if the size of this peak was significant ( $P < 0.05$ ) compared to peaks obtained from the shuffled surrogate histograms (identical to the procedure described earlier in the section Syllable-related neural activity). To eliminate the possibility of including syllable-related activity as bout-related activity, we did not consider a neuron to be exhibiting pre-bout activity if the neuron showed a peak in the bout-onset aligned histogram and a peak at a similar latency (less than 25 ms apart) in the syllable-onset aligned histogram. We considered a neuron to be exhibiting ‘post-bout activity’ if there was a significant peak in the bout-offset aligned histogram (Extended Data Fig. 2j, k) in a 300 ms window after bout-offset.

**Quantification of the rhythmic neural activity.** To quantify the rhythmic neural activity of HVC projection neurons, we used four different methods: inter-burst interval, spike-train autocorrelation, spectrum of the spike train, and cepstrum of the spike train. Only spikes that were produced during singing (that is, between the onset of the first syllable and the offset of the last syllable in the bout) were used for the calculation of these measures. (1) Inter-burst interval. Intervals between burst times were calculated and the peak between 80–1,000 ms was found. (2) Spike-train autocorrelation. To quantify the second-order statistics of the firing pattern of HVC neurons, spike-train autocorrelation, expressed as a conditional firing rate<sup>61</sup>, was calculated, and the peak between 80–1,000 ms was found. The width of the centre peak indicates the width of bursts, and multiple side lobes with regular intervals indicate rhythmic bursting. (3) Spectrum of the spike train. Rhythmicity of the single-unit activity was also quantified in the frequency domain using multi-taper spectral analysis of spike trains treated as point processes<sup>62</sup>. We used the Chronux software to calculate the spectrum for the spike trains<sup>63,64</sup>. First, bouts of singing were segmented into non-overlapping analysis windows of 1.5 s long, and then the spectrum for each window was calculated using multi-taper spectral analysis with time-bandwidth product  $NW = 3/2$  and the number of tapers  $K = 2$ . To obtain the mean spectrum for a given neuron, spectra calculated from all the analysis windows were averaged. Finally, we found the peak in the mean spectrum within the range 2–15 Hz. (4) Cepstrum of the spike train. HVC projection

neurons typically exhibited brief rhythmic bursts with precise inter-burst intervals (Fig. 1b, c). Thus, the spectrum of the spike train tended to have peaks at multiples of the fundamental frequency. To represent these burst trains that have regular intervals in a more compact way, we calculated the cepstrum (a technique commonly used in speech processing to extract the period of glottal pulses) of the spike train, defined as the inverse Fourier transform of the log spectrum<sup>65</sup>, and found the peak in the cepstrum between 80–1,000 ms.

To assess the significance of the peaks in these four measures, we compared the distribution of peak amplitude obtained from the real data with that of the surrogate data obtained by shuffling the bursts times. For this shuffling procedure, we first identified all the bursts during a bout of singing as described above. We then randomly placed bursts sequentially in an interval that has the same duration as the song bout; when spikes from two bursts were closer than 30 ms, we repeated the random placement until they were spaced by more than 30 ms. Note that this randomization procedure only shuffles the burst times and preserves both the number of bursts and the ISIs within bursts. Then, all four metrics listed above were calculated by applying the same method to these surrogate spike trains. This shuffling was repeated (1,000 times for the IBI and autocorrelation, 100 times for the spectrum and cepstrum) and the *P* values of the peak were calculated by analysing the frequency at which the peaks from the surrogate spike trains were larger than the peak obtained from real data. A neuron was considered to exhibit ‘rhythmic’ bursting if it had significant peaks in at least two of the four metrics. The period of the rhythm was defined as the location of the largest peak of spike-train autocorrelation between 80–1,000 ms.

**Quantification of the probabilistic neural activity during the protosyllable stage (Extended Data Fig. 2p).** Although many HVC projection neurons recorded in the juvenile bird exhibited rhythmic bursts, these bursts did not occur reliably on every cycle of the rhythm, but instead participated probabilistically (Fig. 2a). To quantify the degree of participation, we first extracted the protosyllables based on syllable duration (see earlier section Syllable classification and labelling) and examined the fraction of protosyllables in which at least one spike occurred (time-window from 30 ms before protosyllable onset to 10 ms after protosyllable offset). The fraction of protosyllables in which the neuron was active was obtained for all the HVC projection neurons recorded during the protosyllable stage that showed a significant rhythmic bursting (Extended Data Fig. 2p).

**Analysis of simultaneously recorded pairs of neurons (Extended Data Fig. 2q, r).** To test whether probabilistic bursting of neurons in the protosyllable stage is coordinated across many neurons, we analysed the correlation between pairs of simultaneously recorded neurons (Fig. 2a, bottom). This analysis was restricted to pairs of neurons that were rhythmically bursting ( $n = 11$  pairs, 3 birds). Bursting activity of each neuron was converted to a binary string corresponding to its participation in each protosyllable (for the definition of protosyllables, see earlier section Syllable classification and labelling). The activity of a neuron was assigned a ‘1’ for a protosyllable if the neuron exhibited activity in a time-window from 30 ms before protosyllable onset to 10 ms after protosyllable offset, and ‘0’ if it did not. Only activity during protosyllables was analysed to avoid including the highly variable subsong syllables, which are likely generated by circuits outside HVC<sup>27,34</sup>. For simultaneously recorded pairs of neurons, this procedure resulted in two binary strings corresponding to the protosyllable-related activity of each neuron. We then calculated the coefficient of determination  $r^2$  by taking the square of the Pearson’s correlation coefficient  $r$  between the two binary strings. The distribution of coefficient of determination is shown in Extended Data Fig. 2q (median  $r^2 = 0.072$ , 11 pairs).

We also carried out a mutual information analysis to quantify whether the activity of one neuron was predictive of the set of protosyllables for which the other neuron was active. Using the same binary representation described above, we calculated the joint probability distribution describing the four possible states of activity (neither neuron spikes, neuron A spikes, neuron B spikes, both neurons spike). The mutual information was computed from this joint distribution (Extended Data Fig. 2r, median mutual information = 0.056 bits, 11 pairs).

Both the correlation and mutual information were extremely low, suggesting that different projection neurons participated on relatively independent sets of protosyllables. These findings suggest that individual projection neurons participate probabilistically and largely independently in an ongoing rhythmic protosyllable sequence within HVC.

**Analysis of coverage by HVC projection neuron bursts (Extended Data Fig. 2s, t).** We wondered whether projection neuron bursts effectively span the entire duration of juvenile song syllables, or whether bursts are highly localized to specific times, leaving other times in the syllable unrepresented<sup>22</sup>. It is clear from the syllable aligned raster plots that some syllables were completely covered by bursts (for example, Fig. 3h, syllable ‘C’), while other syllables showed some gaps



in the burst coverage (for example, Fig. 4i, syllable 'A'). To further quantify this aspect of the HVC representation during singing, we analysed the fraction of time within the syllables of juvenile birds that were 'covered' by the recorded projection neurons bursts ('covered fraction'). This analysis was restricted to syllables with more than 10 associated bursts.

We first determined the region of the song syllable covered by each HVC projection neuron burst. We generated a histogram of syllable-onset or -offset aligned spike times recorded from a single neuron over every recorded rendition of the song syllable. Initial identification of candidate burst events was determined by smoothing the histogram (9 ms sliding square window, 1 ms steps), and setting a threshold to define a window in which to analyse burst spikes (2 Hz for protosyllable stage birds; 10 Hz threshold for older juveniles). To eliminate low-probability spike events, we only considered bursts for which spiking activity (at least one spike) occurred in the candidate burst window on at least 25% of the renditions for that syllable. Bursts were included only if they occurred between 30 ms before syllable onset and 10 ms after syllable offset.

For candidate bursts that met these criteria, all spikes occurring in the burst window were considered as contributing to that burst. Based on earlier measurements of postsynaptic currents and potentials of HVC and RA neurons<sup>66</sup>, each HVC spike in the burst window was conservatively assumed to exert a postsynaptic effect lasting no more than 5 ms. Thus, each spike in the data set was replaced with a 5 ms postsynaptic square pulse (beginning at the spike time). We considered a region of the syllable to be 'covered' by this burst if at least three of these post-synaptic pulses overlapped at that time within the burst, across renditions of the syllable. This procedure yielded a small 'patch' of time covered by the burst. The patches associated with each different neuron were combined with a logical 'OR' operation to determine the total coverage time of the syllable (again in a window from 30 ms before syllable onset to 10 ms after syllable offset). The covered time was divided by the duration of the syllable window to determine the covered fraction. Only syllables that had more than 10 neurons bursting within the syllable window were analysed. This criterion excluded syllables from bird 3 (shown in Extended Data Fig. 8), from which relatively few neurons were recorded.

While most syllables had nearly complete burst coverage (>90%), one syllable had coverage of only 73% (Extended Data Fig. 2t), which could potentially be due to the relatively smaller number of neurons recorded in this bird. Thus, we asked whether the measured coverage is consistent with sparse sampling of the recorded bursts from a large number of uniformly placed bursts. To simulate this, we calculated the covered fraction for 1,000 surrogate data sets in which the 'covered patches' for each burst were randomly shuffled within the syllable. A random offset was added to the time of each patch, and a circular shift was used, allowing the patches to wrap around the edges of the syllable window. The distribution of covered fractions was determined over all shuffled surrogate data sets, and the 2.5–97.5 percentiles (95% confidence interval) of this distribution were determined (shown as vertical grey bars in Extended Data Fig. 2t). For all syllables, the observed covered fraction was consistent with that expected for random sampling from a uniform underlying distribution of burst times.

**Shared and specific neurons.** To examine whether a given HVC projection neuron was active during multiple syllable types ('shared' neuron) or was active only during a specific syllable type ('specific' neuron), we first constructed a syllable-onset aligned histogram (1 ms bin, smoothed over 20 bins) for each syllable type. Spike times were linearly time warped<sup>67</sup> to the mean duration of that syllable to reduce the trial-to-trial variability in the spike timing associated with the variation in the syllable duration. Next, we found the peak in the firing rate histogram in the interval between 30 ms before syllable onset and 10 ms after syllable offset. We visually inspected the syllable-aligned histograms, and adjusted the interval if necessary to avoid the same burst being detected twice (that is, being associated with an offset of one syllable and an onset of the next syllable). The significance of this peak was determined by comparing it with the peak size obtained from the shuffled histogram using the same method described earlier (in Syllable-related neural activity section).

We defined 'shared' and 'specific' neurons in the context of a particular syllable differentiation process (for example,  $\beta$  and  $\gamma$  from bird 1 in Fig. 3;  $\alpha$  and  $\beta$  from bird 2 in Fig. 4; B and D from bird 1 in Extended Data Fig. 7). 'Specific' neurons were defined as neurons that had a significant peak in the syllable-aligned histogram for only one syllable type, whereas 'shared' neurons were defined as neurons that had significant peaks for both syllable types. We took a conservative approach and only considered a neuron to be shared if the peak was significant for both syllable types. However, some neurons classified as specific had weak activity for the other syllable that did not reach significance (for example, Extended Data Fig. 6f). In other words, we believe this method likely underestimated the fraction of neurons with shared activity.

Our method likely underestimated the incidence of shared neurons for another reason as well. Specifically, we defined shared and specific neurons in the context of a particular pair of syllables undergoing differentiation. For example, in a bird that exhibited hierarchical differentiation (bird 1; Extended Data Fig. 7), we saw examples of neurons that were B-specific when considering B-C differentiation but shared when considering B-D differentiation. Thus, when considering all the syllables in the motif, our definition of shared and specific neuron based on syllable pairs will underestimate the fraction of shared neurons and overestimate the fraction of specific neurons.

**Quantification of the similarity of latencies in shared neurons (Extended Data Fig. 4a–d and Extended Data Fig. 8i, j).** To test whether shared neurons were active at similar latencies for multiple syllable types, we first calculated the latency of the peak in the syllable-onset- or offset-aligned histograms. We then plotted the latency of the peak for one syllable against that of another syllable (Extended Data Fig. 4a–d). When a shared neuron was active for three or more syllables, two syllables associated with two highest firing rates were chosen. To quantify whether shared neurons were active at similar latencies for two syllable types, we calculated the Pearson's correlation coefficient  $r$  between the two latencies across shared neurons, and the  $P$  value under the null hypothesis that  $r = 0$ .

For the bird whose song was segmented based on the phase of the rhythm (bird 3, Extended Data Fig. 8), we asked whether bursts of shared neurons during different syllables occurred at similar phases of the rhythm. To quantify the phase of the neural activity, we first detected the burst times during singing, and for each burst, we assigned an instantaneous phase extracted from the song using the Hilbert transform (see the section on phase segmentation above). Then, the mean phase of all the bursts produced during a particular syllable type was calculated ( $\varphi_i$ , where  $i = 1, 2, \dots, 5$  indicates syllables). Finally, the two syllable types were chosen for which the neuron participated most reliably, and the difference between the mean phases for these two syllables ( $|\Delta\varphi| = |\varphi_m - \varphi_n|$ , where  $m$  and  $n$  are syllable indices) was obtained (Extended Data Fig. 8i). We tested the significance of this value by comparing the value of  $|\Delta\varphi|$  against that obtained from the shuffled data where the pairing of phases were randomized across all shared neurons (Extended Data Fig. 8j; 1,000 shuffles).  $P$  values were obtained by analysing the frequency with which  $|\Delta\varphi|$  of surrogate data was smaller than that of the real data, and  $P < 0.05$  was considered significant.

**Quantification of the activity level difference in shared neurons (Extended Data Fig. 4i, j).** To quantify the difference in the activity level for multiple syllable types in the shared neurons, we calculated the 'bias' defined as follows:

$$\text{Bias} = 1 - \frac{\min(r_1, r_2)}{\max(r_1, r_2)}$$

where  $r_i$  is the peak firing rate in the syllable-aligned histogram for syllable  $i$ . Bias of 0 indicates equal activity level for both syllable types, whereas bias of 1 indicates exclusive activity for only one of the syllable types (Extended Data Fig. 4j).

**Analysis of acoustic features associated with bursts of shared neurons (Extended Data Fig. 5).** We wondered if the bursts of shared neurons were associated with different acoustic signals in the shared syllables at the time of the bursts. (An alternative possibility is that shared neurons burst only at times within the emerging syllable types when the acoustic signals are identical.) An example of a neuron analysed here is shown in Extended Data Fig. 5a (from the same data shown in Fig. 3e). This neuron bursts just after the onset of both syllables  $\beta$  and  $\gamma$ . We analysed the acoustic differences in a 0–50 ms analysis window after the burst time, but were most interested in acoustic differences in a narrower premotor window (10–40 ms), as this corresponds to the premotor latency for which one expects HVC neurons to exert an effect on vocal output<sup>29,58,68</sup>.

For each neuron analysed, all syllables in which the neuron generated a burst were identified. The analysis was carried out for every syllable rendition on which the neuron burst, and was restricted to only those syllables. Syllables had previously been labelled by type (that is,  $\beta$  and  $\gamma$ ). We first directly visualized the spectral differences between the two syllable types using a sparse contour representation<sup>69,70</sup>, which is suitable for constructing an 'average' spectrogram. The analysis was carried out on the sound signal extracted from a 50 ms window after each burst. In many cases, this spectral representation revealed consistent differences between the different syllable types in this analysis window (Extended Data Fig. 5b, c).

One complication is that some of the shared neurons burst before syllable onsets or immediately before syllable offsets such that the 10–40 ms window after the bursts was obscured by silent gaps (9 of 24 HVC<sub>RA</sub> neurons and 59 of

120 HVC<sub>X</sub> neurons were obscured). These neurons were excluded from the analysis of acoustic difference.

We further quantified differences in the acoustic signals by extracting time varying acoustic and spectral features in a window 0–50 ms after burst time (see subsection Definition of bursts). We used 8 acoustic features previously established to analyse birdsongs (Wiener entropy, spectral centre of gravity, spectral width, pitch, pitch goodness, sound amplitude, amplitude modulation, frequency modulation)<sup>51,55</sup>. The 8-dimensional vector of features was calculated in 1 ms steps over the 50 ms analysis window (Extended Data Fig. 5d, e).

Because each syllable was labelled, we could determine if the feature trajectories were significantly different for syllables labelled  $\beta$  and those labelled  $\gamma$ , and make this determination at every time step in the analysis window (Extended Data Fig. 5d, e; s.e.m. indicated by shaded region around mean trajectory). Rather than quantify the difference in these trajectories one feature at a time, we used Fisher's discriminant analysis<sup>71</sup> to project the 8-dimensional acoustic feature vector onto a single dimension that gives maximum separability between the two syllable types. The projected direction is determined independently at each time point, and the feature vectors of all syllable renditions are projected, at each time point, to yield a distribution of projected samples. For most neurons, the different syllable types produce visibly different distributions of projected samples (Extended Data Fig. 5f) indicating distinct acoustic structure. The separability of the distributions (in one dimension) of projected samples for different syllable types was quantified using the  $d'$ -prime metric ( $d'$ ), corresponding to the distance between the means of the distributions, normalized by the pooled variance<sup>70</sup>:

$$d' = \frac{\mu_A - \mu_B}{\sqrt{\frac{1}{2}(\sigma_A^2 + \sigma_B^2)}}$$

Because the features evolve in time, this analysis is carried out independently at each 1 ms step in the 50 ms analysis window, and the  $d'$  was plotted as a function of time (Extended Data Fig. 5g). Statistical significance of the  $d'$  trajectory was assessed by randomizing the syllable labels and rerunning the  $d'$  analysis on shuffled data sets ( $N = 1,000$  shuffles). For each randomization, the peak value of  $d'$  in 10–40 ms premotor window was recorded; significance threshold was set as the 95 percentile of the distribution of these peak values. A shared neuron was determined to have significant acoustic difference between the shared syllables only if the  $d'$  trajectory remained above this significance threshold for the entire premotor window of 10–40 ms after the burst. Note that, in the simulated data, none of the 1,000 surrogate runs generated a  $d'$  trajectory that met this stringent criterion.

**Statistics.** Results are expressed as the mean  $\pm$  s.d. or s.e.m. as indicated. For  $\chi^2$  tests, if the contingency table included a cell that had an expected frequency less than 5, Fisher's exact test was used<sup>72</sup>. All tests were two-sided, and  $P < 0.05$  was considered significant. Bonferroni correction was used to account for multiple comparisons.

**Figure 1f.** The statistical significance of developmental changes in the fraction of HVC neurons that were syllable-aligned was assessed in two different ways: (1) Each stage was compared with the adult stage using the  $\chi^2$  test followed by a post-hoc pairwise test. (2) To quantify the developmental trend in the fraction of syllable-locked neurons, we calculated Pearson's correlation coefficient  $r$  between the binary value for each neuron (0, unlocked; 1, locked) and song stage (subsong: 1, protosyllable; 2, multi-syllables; 3, motif; 4, adult; 5). The  $P$  value was calculated under the null hypothesis that  $r = 0$ . The significance of the developmental trend for rhythmic bursting was calculated similarly. Similar results were obtained for correlation between these metrics and the age at which each neuron was recorded, rather than song stage.

**Figure 1g.** The statistical significance of developmental changes in the period of the HVC rhythm was also assessed in two different ways: (1) Each song stage was compared with the adult stage using the Kruskal–Wallis test followed by a post-hoc pairwise test. (2) To quantify the developmental trend in the period of the HVC rhythm, we calculated Pearson's correlation coefficient  $r$  between burst period and song stage. Similar results were obtained for correlation between burst period and the age at which each neuron was recorded.

**Figure 2c.** The Wilcoxon rank-sum test was used to test whether the median of the syllable-onset aligned latency distribution was different between subsong and protosyllable stages.

**Figures 3g, h and 4h, i.** To test whether the fraction shared neurons differed between early and late stages of syllable differentiation, we used the  $\chi^2$  test on a  $2 \times 2$  contingency table (shared/specific, early/late). Regarding across all birds, to calculate whether the fraction of shared neurons differed between early and late stages of syllable differentiation over all birds ( $n = 5$  syllable pairs

in 3 birds), we used the Cochran–Mantel–Haenszel test for repeated tests of independence<sup>73</sup>.

**Extended Data Fig. 1a.** To quantify the relation between song stage and age, we calculated Spearman's rank correlation coefficient  $\rho$  and the  $P$  value under the null hypothesis that  $\rho = 0$ .

**Extended Data Fig. 1c.** We computed the statistical significance of developmental changes in burst width (top) and firing rate during bursts (bottom) by using the Kruskal–Wallis test followed by a post-hoc pairwise test to compare each stage with the adult stage.

**Extended Data Fig. 2m–o.** To test whether fraction of syllable-locked neurons (Extended Data Fig. 2m), fraction of rhythmic neurons (Extended Data Fig. 2n), and period of HVC rhythm (Extended Data Fig. 2o) significantly differed between HVC<sub>RA</sub> and HVC<sub>X</sub>, we used  $\chi^2$  test for all the pairwise comparisons with Bonferroni correction for multiple comparisons.

**Extended Data Fig. 4a–d.** To calculate the relation between latencies of bursts associated with shared neurons, we calculated the Pearson's correlation coefficient  $r$  together with the  $P$  value under the null hypothesis that  $r = 0$ .

**Extended Data Fig. 5m, n.** To test whether the mean  $d'$  metric was different between HVC<sub>RA</sub> and HVC<sub>X</sub>, we used the Wilcoxon rank-sum test. Only neurons with  $d'$  trajectories that were significant (continuously from 10–40 ms) were included in this comparison.

**Neural model of chain formation and splitting.** Code used to simulate the model is available as Supplementary Information. To illustrate a potential mechanism of chain splitting, we chose to implement the model as simply as possible. We modelled neurons as binary units and simulated their activity in discrete time steps<sup>44</sup>; at each time step (10 ms), the  $i$ th neuron either bursts ( $x_i = 1$ ) or is silent ( $x_i = 0$ ).

**Network architecture.** A network of 100 binary neurons is recurrently connected in an all-to-all manner, with  $W_{ij}$  representing the synaptic strength from presynaptic neuron  $j$  to postsynaptic neuron  $i$ . Self-excitation is prevented by setting  $W_{ij} = 0$  for all  $i$  at all times<sup>44</sup>. During learning, the strength of each synapse is constrained to be within the interval  $[0, w_{\max}]$ , while the total incoming and outgoing weights of each neuron are both constrained by the “soft bound”  $W_{\max} = m^* w_{\max}$  where  $m$  represents a target number of saturated synapses per neuron<sup>44</sup> (see section Synaptic plasticity rule for details). Note that  $w_{\max}$  represents a hard maximum weight of each individual synapse, while  $W_{\max}$  represents a soft maximum total synaptic input or output of any one neuron. Synaptic weights are initialized with random uniform distribution such that each neuron receives, on average, its maximum allowable total input,  $W_{\max}$ .

**Network dynamics.** The activity of each neuron in the network was determined in two steps: calculating the net feedforward input that comes from the previous time step; then determining whether that is enough to overcome the recurrent inhibition in the current time step.

First, the net feedforward input to the  $i$ th neuron at time step  $t$ ,  $A_i^{\text{net}}(t)$ , was calculated by summing the excitation, feedforward inhibition, neural adaptation, and external inputs:

$$A_i^{\text{net}}(t) = [A_i^{\text{E}}(t) - A_i^{\text{Iff}}(t) - A_i^{\text{adapt}}(t) + B_i(t) - \theta_i]_+$$

where  $[z]_+$  indicates a rectification (equal to  $z$  if  $z > 0$  and 0 otherwise).  $A_i^{\text{E}}(t) = \sum_j W_{ij} x_j(t-1)$  is the excitatory input from network activity on the previous time step.  $A_i^{\text{Iff}}(t) = \beta \sum_j x_j(t-1)$  is a global feedforward inhibitory input<sup>44</sup>, where  $\beta$  sets the strength of this feedforward inhibition.  $A_i^{\text{adapt}}(t) = \alpha y_i$  is an adaptation term<sup>44</sup> where  $\alpha$  is the strength of adaptation, and  $y_i$  is a low-pass filtered record of recent activity in  $x_i$  with time constant  $\tau_{\text{adapt}} = 40$  ms; that is  $\tau_{\text{adapt}} \frac{dy_i}{dt} = -y_i + x_i$ ;  $B_i(t)$  is the external input to neuron  $i$  at time  $t$ . For seed neurons, this term consists of training inputs (see section on Seed neurons). For non-seed neurons, it consists of random inputs with probability  $p_{\text{in}} = 0.01$  in each time step and size  $W_{\max}/10$ . Finally,  $\theta_i$  is a threshold term used to reduce the excitability of seed neurons, making them less responsive to recurrent input than are other neurons in the network. For seed neurons,  $\theta_i = 10$  and for non-seed neurons,  $\theta_i = 0$ . Including this term improves robustness of the training procedure by eliminating occasional situations in which seed neuron activity may be dominated by recurrent rather than external inputs. In these cases, external inputs may fail to exert proper control of network activity.

Second, we determined whether the  $i$ th neuron will burst or not at time step  $t$  by examining whether the net feedforward input,  $A_i^{\text{net}}(t)$ , exceeds the recurrent inhibition,  $A_i^{\text{L-rec}}(t)$ . We implemented recurrent inhibition by estimating the total input to the network at time  $t$ :

$$A_i^{\text{L-rec}}(t) = \gamma \sum_j A_j^{\text{net}}(t)$$



and feeding it back to all the neurons. Parameter  $\gamma$  sets the strength of the recurrent inhibition. We assume that this recurrent inhibition operates on a fast time scale<sup>48</sup> (that is, faster than the duration of a burst). Thus, the final output of the  $i$ th neuron at time  $t$  becomes:

$$x_i(t) = \Theta[A_i^{\text{net}}(t) - A_i^{\text{rec}}(t)]$$

where  $\Theta[z]$  is the Heaviside step function (equal to 1 if  $z > 0$  and 0 otherwise). To induce splitting,  $\gamma$  was gradually stepped up to  $\gamma_{\text{split}}$  following a sigmoid with time constant  $\tau_\gamma$  and inflection point  $t_0$ :

$$\gamma(t) = \frac{\gamma_{\text{split}}}{1 + e^{-(t-t_0)/\tau_\gamma}}$$

**Seed neurons.** A subset of neurons was designated as seed neurons, which received external training inputs used to shape network activity during learning<sup>43,45</sup>. The external training inputs activate seed neurons at syllable onsets, reflecting the observed onset-related bursts of HVC neurons during the subsong stage (Fig. 1a). The pattern of these inputs was adjusted in different stages of learning, and each strategy of syllable learning was implemented by different patterns of seed neuron training inputs.

**Alternating differentiation (Fig. 5a–e).** Ten neurons were designated as seed neurons and received strong external input ( $W_{\text{max}}$ ) to drive network activity. In the subsong stage, seed neurons were driven (by external inputs) synchronously and randomly with probability 0.1 in each time step corresponding to the random occurrence of syllable onsets in subsong<sup>27,34</sup>. This was done only to visualize network activity; no learning was implemented at the subsong stage. During the protosyllable stage, seed neurons were driven synchronously and rhythmically with a period  $T = 100$  ms. The protosyllable stage consisted of 500 iterations of 10 pulses each. To initiate chain splitting, the seed neurons were divided into two groups and each group was driven on alternate cycles. The splitting stage consisted of 2,000 iterations of 5 pulses in each group of seed neurons (1 s total per iteration, as in the protosyllable stage).

**Motif strategy (Extended Data Fig. 10e–h).** This was implemented in a similar manner as alternating differentiation, except that 9 seed neurons were used, and for the splitting stage, seed neurons were divided into 3 groups of 3 neurons, each driven on every third cycle.

**Bout-onset differentiation (Extended Data Fig. 10a–d).** Seed neurons were divided into two groups: 5 bout-onset seed neurons and 5 protosyllable seed neurons. At all learning stages, external inputs were organized into bouts consisting of four separate input pulses, and bout-onset seed neurons were driven at the beginning of each bout. Then, 30 ms later, protosyllable seed neurons were driven three times with an interval of  $T = 100$  ms. In the protosyllable stage, inputs to all seed neurons were of strength  $W_{\text{max}}$ . In the splitting stage, the input to protosyllable seed neurons was decreased to  $W_{\text{max}}/10$ . This allowed neurons in the bout-onset chain to suppress, through fast recurrent inhibition, the activity of protosyllable seed neurons during bout-onset syllables.

Each iteration of the simulation was 5 s long, consisting of 10 bouts, described directly above, with random inter-bout intervals. The protosyllable stage consisted of 100 iterations, and the splitting stage consisted of 500 iterations.

**Bout-onset syllable formation (Extended Data Fig. 10i–k).** Input to seed neurons was set high ( $2.5 \cdot W_{\text{max}}$ ), and maintained at this high level throughout development. This prevented protosyllable seed neurons from being inhibited by neurons in the bout-onset chain. Furthermore, strong external input to the protosyllable seed neurons terminated activity in the bout-onset chain through fast recurrent inhibition, thus preventing further growth of the bout-onset chain, as occurs in bout-onset differentiation.

As in bout-onset differentiation, each iteration of the simulation was 5 s long, consisting of 10 bouts with random inter-bout intervals. The protosyllable stage consisted of 100 iterations, and the splitting stage consisted of 500 iterations.

**Synaptic plasticity rules.** As in previous models<sup>43,44</sup>, we hypothesized two plasticity rules in our model: Hebbian spike-timing dependent plasticity (STDP) to drive sequence formation<sup>74,75</sup>, and heterosynaptic long term depression (hLTD) to introduce competition between synapses of a given neuron<sup>43,44</sup>. STDP is governed by the antisymmetric plasticity rule with a short temporal window (one burst duration):

$$\Delta_{ij}^{\text{STDP}}(t) = \eta [x_i(t)x_j(t-1) - x_i(t-1)x_j(t)]$$

where the constant  $\eta$  sets the learning rate. hLTD limits the total strength of weights for neuron  $i$ , and the summed weight limit rule for incoming weights is given by:

$$\Delta_{i^*}^{\text{hLTD}}(t) = \eta \left[ \sum_k (W_{ik}(t-1) + \Delta_{ik}^{\text{STDP}}(t)) - W_{\text{max}} \right]_+$$

and for outgoing weights from neuron  $j$ :

$$\Delta_{ij}^{\text{hLTD}}(t) = \eta \left[ \sum_k (W_{kj}(t-1) + \Delta_{kj}^{\text{STDP}}(t)) - W_{\text{max}} \right]_+$$

At each time step, total change in synapse weight is given by the combination of STDP and hLTD:

$$\Delta W_{ij}(t) = \Delta_{ij}^{\text{STDP}}(t) - \varepsilon \Delta_{i^*}^{\text{hLTD}}(t) - \varepsilon \Delta_j^{\text{hLTD}}(t)$$

where  $\varepsilon$  sets the relative strength of hLTD.

**Model parameters: subsong (Fig. 5a).** In our implementation of the subsong stage, there was no learning. Subsong model parameters were:  $\beta = 0.115$ ,  $\alpha = 30$ ,  $\eta = 0$ ,  $\varepsilon = 0$ ,  $\gamma = 0.01$ .

**Model parameters: alternating differentiation (Fig. 5b–d).** After subsong, learning progressed in two stages: the protosyllable stage and the splitting stage. Parameters that remained constant over development were:  $\beta = 0.115$ ,  $\alpha = 30$ ,  $\eta = 0.025$ ,  $\varepsilon = 0.2$ . To induce chain splitting,  $w_{\text{max}}$ , the maximum allowed strength of any synapse, was increased from 1 to 2,  $m$  was decreased from 10 to 5, and  $\gamma$  was increased from 0.01 to 0.18 following a sigmoid with time constant  $\tau_\gamma = 200$  iterations and inflection point  $t_0 = 500$  iterations into the splitting stage. No change in parameters occurred before the chain-splitting stage.

**Model parameters: bout-onset differentiation (Extended Data Fig. 10a–d).** Parameters that remained constant over development were:  $\beta = 0.13$ ,  $\alpha = 30$ ,  $\eta = 0.05$ ,  $\varepsilon = 0.14$ . To induce chain splitting,  $w_{\text{max}}$  was increased from 1 to 2,  $m$  was decreased from 5 to 2.5, and  $\gamma$  was increased from 0.01 to 0.04 following a sigmoid with time constant  $\tau_\gamma = 200$  iterations and inflection point  $t_0 = 250$  iterations into the splitting stage.

**Model parameters: motif strategy (Extended Data Fig. 10e–h).** Parameters that remained constant over development were:  $\beta = 0.115$ ,  $\alpha = 30$ ,  $\eta = 0.025$ ,  $\varepsilon = 0.2$ . To induce chain splitting,  $w_{\text{max}}$  was increased from 1 to 2,  $m$  was decreased from 9 to 3, and  $\gamma$  was increased from 0.01 to 0.18 following a sigmoid with time constant  $\tau_\gamma = 200$  iterations and inflection point  $t_0 = 500$  iterations into the splitting stage.

**Model parameters: formation of a new syllable at bout onset (Extended Data Fig. 10i–k).** Parameters that remained constant over development were:  $\beta = 0.13$ ,  $\alpha = 30$ ,  $\eta = 0.05$ ,  $\varepsilon = 0.15$ . To induce chain splitting,  $w_{\text{max}}$  was increased from 1 to 2,  $m$  was decreased from 5 to 2.5, and  $\gamma$  was increased from 0.01 to 0.05 following a sigmoid with time constant  $\tau_\gamma = 200$  iterations and inflection point  $t_0 = 250$  iterations into the splitting stage.

**Shared and specific neurons.** Neurons were classified as participating in a syllable type if the syllable onset-aligned histogram exhibited a peak that passed a threshold criterion. The criteria were chosen to include neurons where the histogram peak exceeded 90% of surrogate histogram peaks. Surrogate histograms were generated by placing one burst at a random latency in each syllable. (For example, in the protosyllable stage, the above criterion was found to be equivalent to having 5 bursts at the same latency in a bout of 10 protosyllables.) During the splitting phase, neurons were classified as shared if they participated in both syllable types, and specific if they participated in only one syllable type.

**Visualizing network activity.** We visualized network activity in two ways: network diagrams, and raster plots of population activity (for example, Fig. 5a–d top and bottom panels, respectively). In both cases, we only included neurons that participated in at least one of the syllable types (see earlier section Shared and specific neurons for participation criteria).

**Network diagrams.** Neurons are sorted along the  $x$  axis based on their relative latencies. Neurons are sorted along the  $y$  axis based on the relative strength of their synaptic input from specific neurons (or seed neurons) of each type (red or blue). Lines between neurons correspond to feedforward synaptic weights, and darker lines indicate stronger synaptic weights. For clarity of plotting, only the strongest six outgoing and strongest nine incoming weights are plotted for each neuron.

**Population raster plots.** Neurons are sorted from top to bottom according to their latency. Groups of seed neurons are indicated by magenta arrows. Shared neurons are plotted at the top and specific neurons are plotted below. As for network diagrams, neurons that did not reliably participate in at least one syllable type were excluded.

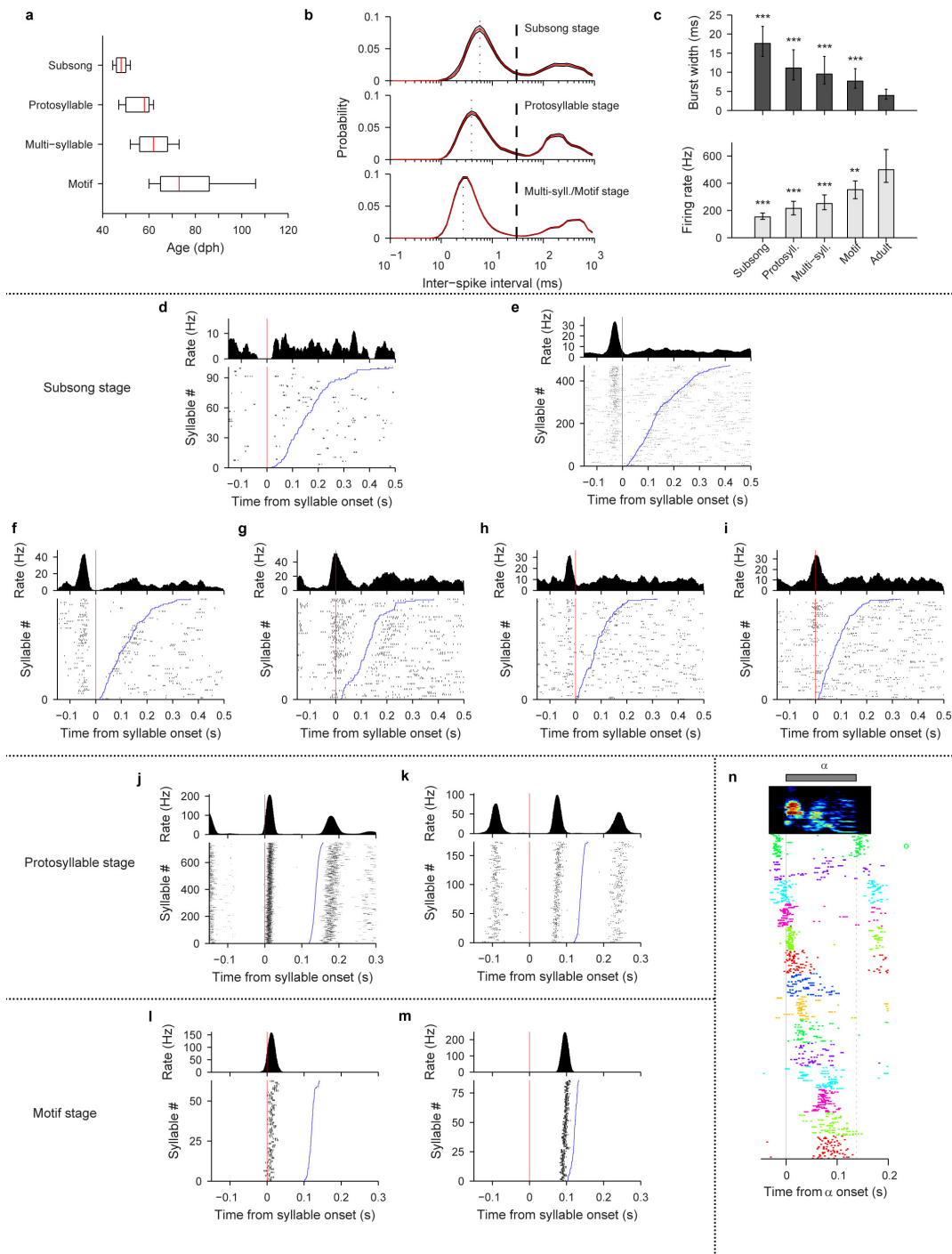
**Further details for Fig. 5a–d.** Panels show network diagrams and raster plots at four different stages. Figure 5a shows subsong stage (before learning), Fig. 5b shows end of protosyllable stage (iteration 500), Fig. 5c shows early chain splitting stage (iteration 992), Fig. 5d shows late chain-splitting stage (iteration 2,500).

**Further details for Extended Data Fig. 10a–d.** Extended Data Fig. 10a shows early protosyllable stage (iteration 5), Extended Data Fig. 10b shows late protosyllable

stage (iteration 100), Extended Data Fig. 10c shows early chain splitting stage (iteration 130), Extended Data Fig. 10d shows late chain splitting stage (iteration 600). **Code availability.** Code used to simulate the model is available as Supplementary Information.

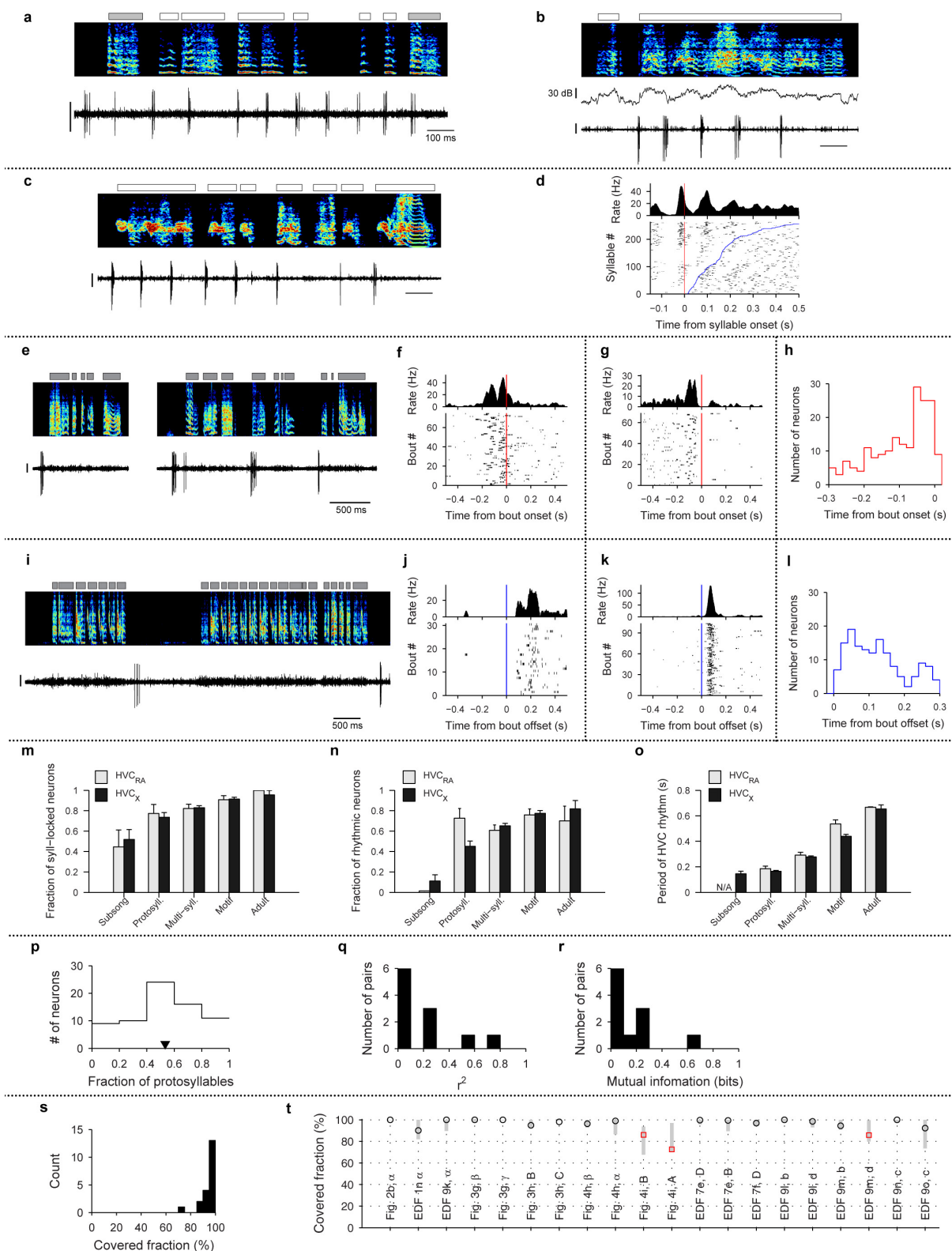
51. Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B. & Mitra, P. P. A procedure for an automated measurement of song similarity. *Anim. Behav.* **59**, 1167–1176 (2000).
52. Tchernichovski, O., Lints, T. J., Deregnaucourt, S., Cimenser, A. & Mitra, P. P. Studying the song development process: rationale and methods. *Ann. NY Acad. Sci.* **1016**, 348–363 (2004).
53. Goller, F. & Daley, M. A. Novel motor gestures for phonation during inspiration enhance the acoustic complexity of birdsong. *Proc. R. Soc. Lond. B* **268**, 2301–2305 (2001).
54. Rajan, R. & Doupe, A. J. Behavioral and neural signatures of readiness to initiate a learned motor sequence. *Curr. Biol.* **23**, 87–93 (2013).
55. Mandelblat-Cerf, Y. & Fee, M. S. An automated procedure for evaluating song imitation. *PLoS One* **9**, e96484 (2014).
56. Fee, M. S. & Leonardo, A. Miniature motorized microdrive and commutator system for chronic neural recording in small animals. *J. Neurosci. Methods* **112**, 83–94 (2001).
57. Okubo, T. S., Mackevicius, E. L. & Fee, M. S. In vivo recording of single-unit activity during singing in zebra finches. *Cold Spring Harb. Protoc.* **2014**, 1273–1283 (2014).
58. Fee, M. S., Kozhevnikov, A. A. & Hahnloser, R. H. Neural mechanisms of vocal sequence generation in the songbird. *Ann. NY Acad. Sci.* **1016**, 153–170 (2004).
59. Hahnloser, R. H., Kozhevnikov, A. A. & Fee, M. S. Sleep-related neural activity in a premotor and a basal-ganglia pathway of the songbird. *J. Neurophysiol.* **96**, 794–812 (2006).
60. Goldberg, J. H. & Fee, M. S. A cortical motor nucleus drives the basal ganglia-recipient thalamus in singing birds. *Nature Neurosci.* **15**, 620–627 (2012).
61. Rieke, F. *Spikes: Exploring the Neural Code* (MIT Press, 1997).
62. Jarvis, M. R. & Mitra, P. P. Sampling properties of the spectrum and coherency of sequences of action potentials. *Neural Comput.* **13**, 717–749 (2001).
63. Bokil, H., Andrews, P., Kulkarni, J. E., Mehta, S. & Mitra, P. P. Chronux: a platform for analyzing neural signals. *J. Neurosci. Methods* **192**, 146–151 (2010).
64. Mitra, P. & Bokil, H. *Observed Brain Dynamics* (Oxford Univ. Press, 2008).
65. Oppenheim, A. V. & Schaffer, R. W. From frequency to quefrency: a history of the Cepstrum. *IEEE Signal Process. Mag.* **21**, 95–106 (2004).
66. Garst-Orozco, J., Babadi, B. & Ölveczky, B. P. A neural circuit mechanism for regulating vocal variability during song learning in zebra finches. *eLife* **3**, e03697 (2014).
67. Leonardo, A. & Fee, M. S. Ensemble coding of vocal control in birdsong. *J. Neurosci.* **25**, 652–661 (2005).
68. Ashmore, R. C., Wild, J. M. & Schmidt, M. F. Brainstem and forebrain contributions to the generation of learned motor behaviors for song. *J. Neurosci.* **25**, 8543–8554 (2005).
69. Lim, Y., Shinn-Cunningham, B. & Gardner, T. J. Sparse contour representations of sound. *IEEE Signal Process. Lett.* **19**, 684–687 (2012).
70. Markowitz, J. E., Ivie, E., Kligler, L. & Gardner, T. J. Long-range order in canary song. *PLOS Comput. Biol.* **9**, e1003052 (2013).
71. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification* 2nd edn (Wiley, 2001).
72. Kanji, G. K. *100 Statistical Tests* 3rd edn (Sage Publications, 2006).
73. McDonald, J. H. *Handbook of Biological Statistics* 3rd edn (Sparky House Publishing, 2014).
74. Abbott, L. F. & Blum, K. I. Functional significance of long-term potentiation for sequence learning and prediction. *Cereb. Cortex* **6**, 406–416 (1996).
75. Dan, Y. & Poo, M. M. Spike timing-dependent plasticity: from synapse to perception. *Physiol. Rev.* **86**, 1033–1048 (2006).
76. Fee, M. S. & Goldberg, J. H. A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. *Neuroscience* **198**, 152–170 (2011).
77. Fiete, I. R., Hahnloser, R. H., Fee, M. S. & Seung, H. S. Temporal sparseness of the premotor drive is important for rapid learning in a neural network model of birdsong. *J. Neurophysiol.* **92**, 2274–2282 (2004).
78. Charlesworth, J. D., Turner, E. C., Warren, T. L. & Brainard, M. S. Learning the microstructure of successful behavior. *Nature Neurosci.* **14**, 373–380 (2011).
79. Ravbar, P., Lipkind, D., Parra, L. C. & Tchernichovski, O. Vocal exploration is locally regulated during song learning. *J. Neurosci.* **32**, 3422–3432 (2012).
80. Walton, C., Pariser, E. & Nottebohm, F. The zebra finch paradox: song is little changed, but number of neurons doubles. *J. Neurosci.* **32**, 761–774 (2012).





**Extended Data Figure 1 | Bursting and syllable-locked activity in HVC projection neurons of juvenile birds.** **a**, Range of bird ages at which songs were classified at different developmental stages (Spearman's rank correlation between age and stage  $\rho = 0.61$ ; red line indicates the median, box indicates the 25–75 percentile, and whiskers indicate 10–90 percentile;  $n = 12, 13, 18$  and  $6$  birds, respectively;  $n = 39, 135, 565$  and  $378$  neurons, respectively). **b**, Interspike-interval (ISI) distributions (mean  $\pm$  s.e.m.) of HVC projection neurons that exhibited spiking during singing, at three stages of vocal development ( $n = 38, 130, 922$  neurons). ISI distributions computed with logarithmic binning show bimodal structure: the peak around 3–5 ms indicates inter-spike intervals within bursts, and a broader peak around 100–400 ms indicates intervals between bursts (dashed line indicates the 30 ms threshold used for defining a burst; dotted line indicates peak). Note the refractory period below 1 ms. **c**, Burst width (top) and firing rate during bursts (bottom) as a function of developmental stage (median  $\pm$  quartiles;  $n = 39, 135, 565, 378$  and  $32$  neurons, respectively;  $**P < 0.01$ ,  $***P < 0.001$  post-hoc comparison with

adult stage). **d–i**, Syllable-onset-aligned raster plots and histograms for neurons recorded during the subsong stage. Syllables are sorted from bottom to top by increasing syllable duration (blue lines indicate syllable offset). **d**, Neuron that did not exhibit significant locking to subsong syllable onsets (RA-projecting neuron, HVC<sub>RA</sub>; 50 dph; bird 7). **e**, Another neuron in the same bird (same neuron as in Fig. 1a; HVC<sub>RA</sub>; 51 dph). **f, g**, Two projection neurons recorded in a different subsong bird (both X-projecting neurons, HVC<sub>X</sub>; 47 and 48 dph, respectively; bird 9). Note different latencies of bursting. **h, i**, Two projection neurons recorded in a different subsong bird (both HVC<sub>X</sub>; 47 and 44 dph, respectively; bird 10). **j**, Syllable-onset-aligned raster plots and histograms showing strong locking to protosyllables (bird 2). **j**, For the same neuron as in Fig. 1b (HVC<sub>RA</sub>; 62 dph). **k**, For another neuron (HVC<sub>RA</sub>; 65 dph). **l, m**, Two neurons recorded in the motif stage (bird 8). **l**, Neuron locked just after syllable onset (HVC<sub>X</sub> neuron; 61 dph). **m**, Same neuron as in Fig. 1c (HVC<sub>RA</sub>; 68 dph) showing locking late in the song syllable. **n**, Population raster of 14 neurons, aligned to protosyllable onsets (56–59 dph; bird 1).

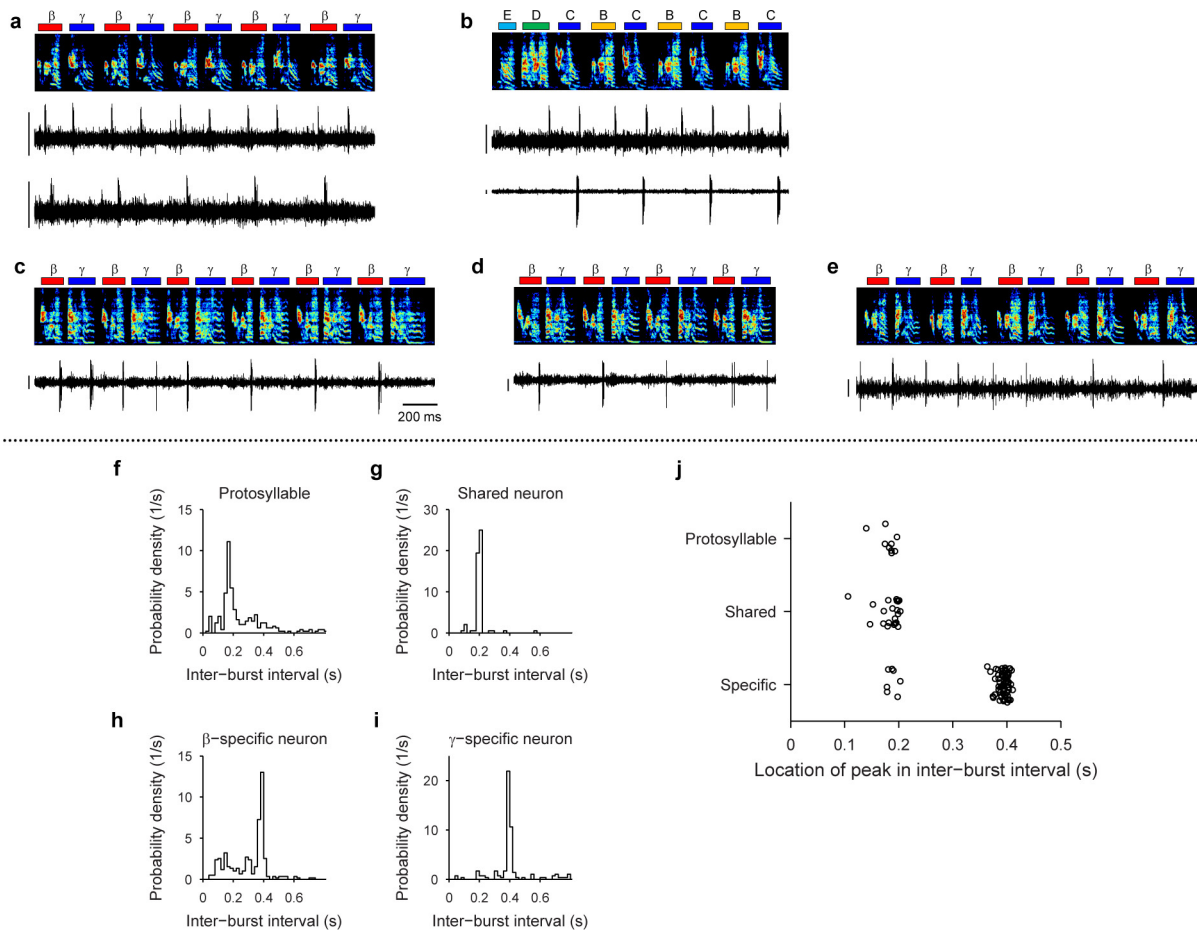


Extended Data Figure 2 | See next page for caption.



**Extended Data Figure 2 | Further analysis and examples of HVC projection neuron activity.** **a–d**, Examples of HVC projection neurons showing rhythmic activity during non-rhythmic song. **a**, Bird 2, HVC<sub>RA</sub> neuron, 57 dph. **b**, Bird 12, HVC<sub>X</sub>, 53 dph. **c**, Bird 12, HVC<sub>RA</sub>, 57 dph. **d**, Syllable onset-aligned raster plot for neuron shown in **c**. Syllables are sorted in order of increasing duration (bottom to top; blue line indicates syllable offset). Also shown (top) is the onset-aligned spike histogram. Note multiple rhythmic bursts during long syllables. Scale bars: panels **a–c**, 1 mV, 100 ms. **e–l**, Bout-related activity of HVC projection neurons. **e**, Bout-onset neuron (HVC<sub>X</sub>; 44 dph; bird 11). **f**, Bout-onset aligned histogram and raster plot for the neuron shown in panel **e**. **g**, Bout-onset aligned histogram and raster plot for the neuron shown in Fig. 1d. **h**, Distribution of pre-bout-onset latencies for all bout-onset neurons ( $n = 187$  neurons, 32 birds). **i**, Bout-offset neuron (HVC<sub>X</sub>; 61 dph; bird 1). **j**, Bout-offset aligned histogram and raster plot for the neuron shown in panel **i**. **k**, Bout-offset aligned histogram and raster plot for the neuron shown in Fig. 1e. **l**, Distribution of post-bout-offset latencies for all bout-offset neurons ( $n = 149$  neurons, 32 birds). Vertical scale bars in panels **e** and **i**, 0.5 mV. **m–o**, Developmental progression of HVC activity analysed separately for HVC<sub>RA</sub> and HVC<sub>X</sub> neurons. **m**, Fraction of neurons temporally locked to syllables (mean  $\pm$  s.e.m.; HVC<sub>RA</sub>: 9, 22, 83, 54 and 10 neurons analysed at each stage, respectively; HVC<sub>X</sub>: 27, 91, 376, 244 and 22 neurons analysed at each stage, respectively). **n**, Fraction of neurons that exhibited rhythmic bursts (HVC<sub>RA</sub>: 9, 22, 83, 54 and 10 neurons, respectively; HVC<sub>X</sub>: 27, 91, 376, 244 and 22 neurons, respectively). **o**, Mean period of HVC rhythmicity as a function of song stage (HVC<sub>RA</sub>: 0, 16, 50, 41 and 7 neurons, respectively; HVC<sub>X</sub>: 3, 41, 245, 189, 18 neurons, respectively). Of the 14 comparisons between HVC<sub>RA</sub> and HVC<sub>X</sub> neurons shown in panels **m–o**, only the period of HVC rhythm (panel **o**) during the motif stage showed significant difference between the cell types ( $P < 0.05$  with Bonferroni correction). **p–r**, Analysis of probabilistic participation in rhythmic activity during protosyllables. **p**, Distribution of the fraction of protosyllables on which spiking occurred ( $n = 70$  neurons). In contrast to the highly reliable bursting of HVC projection neurons in adult birds<sup>19–22</sup>, we found that neurons in the protosyllable stage participated probabilistically (mean: 53% of protosyllables; triangle symbol). **q**, Histogram of the coefficient of determination  $r^2$  for protosyllable

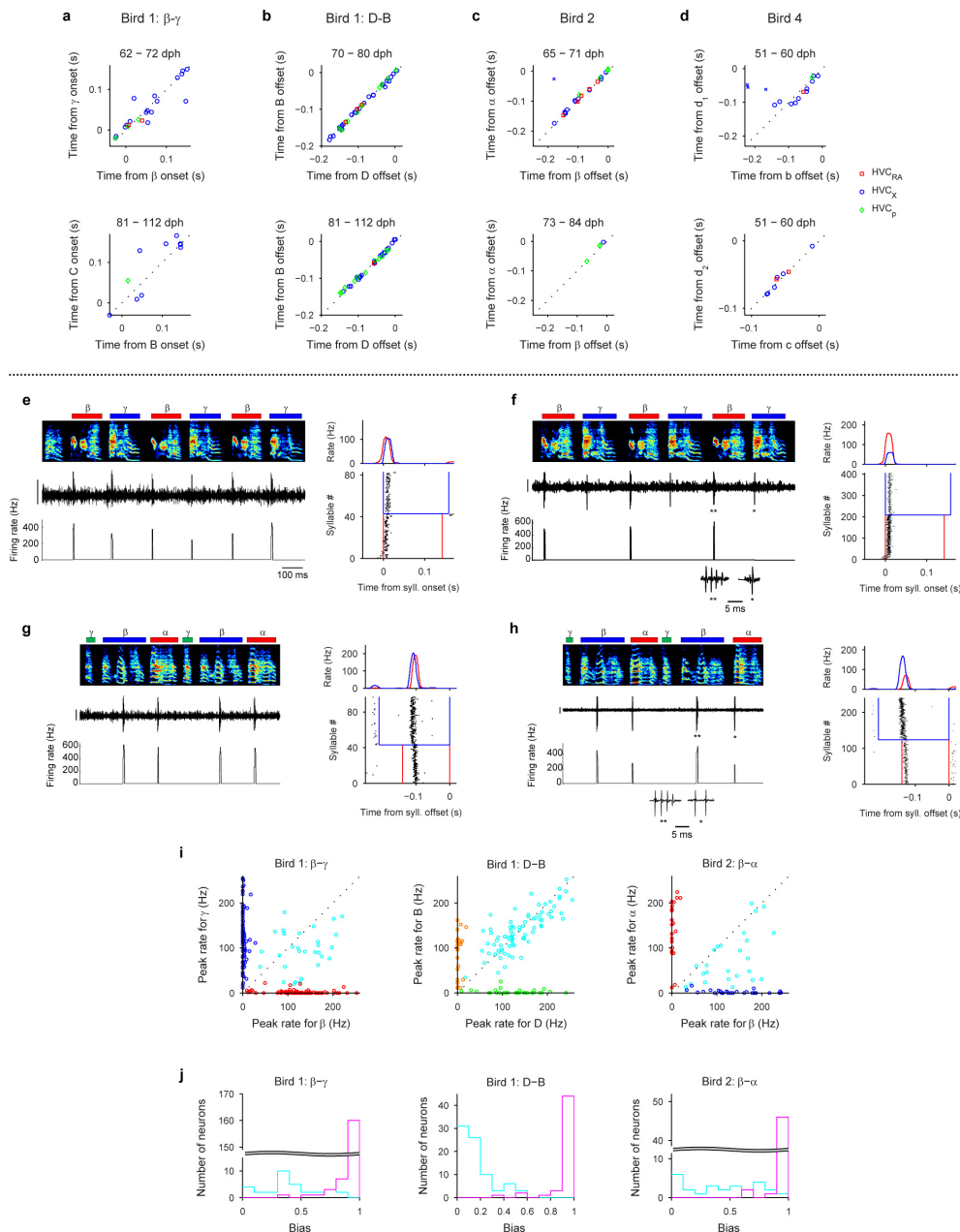
participation across simultaneously recorded pairs of neurons (median  $r^2 = 0.072$ ;  $n = 11$  pairs; see Methods). **r**, Histogram of mutual information for protosyllable participation across simultaneously recorded pairs of neurons (median 0.056 bits;  $n = 11$  pairs; see Methods). **s**, **t**, Analysis of burst coverage by HVC projection neuron bursts. **s**, Summary histogram of the covered fraction for all analysed syllables ( $n = 20$  syllables, 4 birds). Note that 17/20 syllables had a covered fraction higher than 90%. **t**, Covered fraction analysed for 20 syllables for which raster plots are shown in the main or Extended Data figures. Vertical grey bars indicate 95% confidence interval (2.5–97.5 percentile) of coverage expected for random uniform shuffling of the observed bursts (see Methods). Note that for all syllables, the observed coverage is within the confidence interval for randomly shuffled bursts. These findings suggest that, even for the three syllables with coverage less than 90% (indicated with red square symbol), the lower coverage was consistent with undersampling due to the smaller number of recorded neurons in these birds. Regarding two models of HVC coding: our findings bear on several recent models of song representation in HVC. One earlier model hypothesizes that HVC bursts provide timing signals to drive premotor activity<sup>19,58,67</sup> and to control the temporal precision of learning<sup>76–79</sup>. This model implies a continuous, though not necessarily uniform, coverage of HVC bursts throughout song, as observed in our data. Overall, given the very large number of HVC neurons in each hemisphere<sup>80</sup> ( $> 10^4$ ), our measurements are consistent with a continuous representation of timing signals throughout song syllables. Another model of HVC coding has emphasized the finding that bursts may occur more often at particular times in the song, related to ‘gestures’ in the vocal control parameters<sup>22</sup>. Our finding that bursts are more concentrated around syllable onsets early in vocal development suggests that HVC may generate protosyllables as primitive gestures that serve as a scaffold on which later song syllables develop<sup>33</sup>. During development, HVC activity appears to evolve such that, as a population, bursts occur more uniformly throughout song syllables (Fig. 2c), while the activity of individual neurons becomes sparser and more precise. At the same time, one might imagine that vocal gestures become more complex and precise as syllables develop into their adult forms. In this view, the emergence of sequential activity in HVC may be viewed to drive an increasingly complex sequence of gestures.



**Extended Data Figure 3 | Increase in the period of HVC rhythmicity during alternating syllable differentiation.** All data are from bird 1. **a**, Paired recording of a shared neuron (top; HVC<sub>RA</sub>) and a  $\beta$ -specific neuron (bottom; HVC<sub>X</sub>; 69 dph). **b**, Paired recording of a shared neuron (top; HVC<sub>X</sub>) and a C-specific neuron (bottom; HVC<sub>X</sub>; 110 dph). **c**, Neuron switching between shared and specific spiking (HVC<sub>X</sub>; 63 dph). **d**, Same neuron as in **c**, switching from specific to shared spiking. **e**, A different neuron switching from shared to specific spiking (HVC<sub>p</sub>; 68 dph). Scale bars in panels **a–e**, 0.5 mV, 200 ms. **f–i**, Inter-burst interval (IBI) distributions for shared and specific neurons. **f**, For the neuron in Fig. 3c

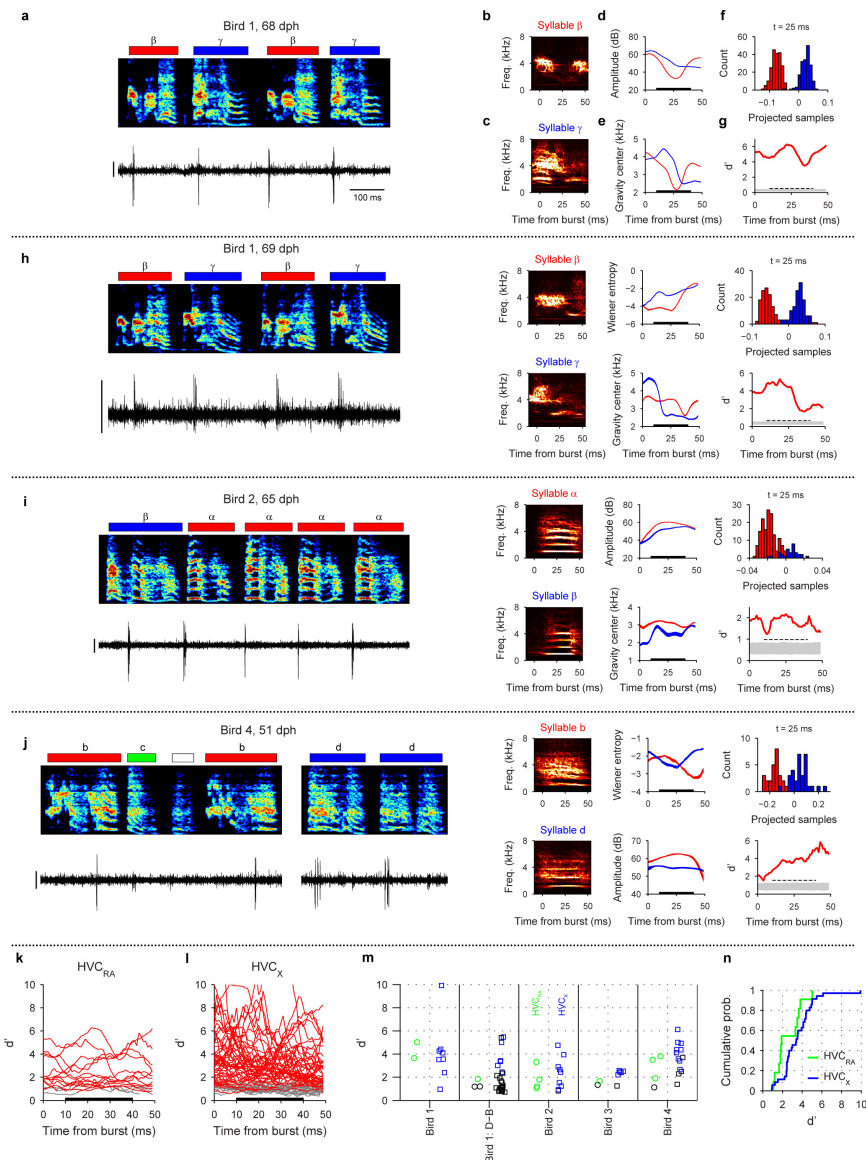
recorded during protosyllable stage. **g**, For the shared neuron shown in the top panel of Fig. 3f. **h**, For the  $\beta$ -specific neuron shown in Fig. 3d. **i**, For a  $\gamma$ -specific neuron (not shown). **j**, Population summary of the 'most-probable IBI' for the neurons recorded during the protosyllable stage ( $n = 9$ ), and during the emergence of syllables  $\beta$  and  $\gamma$  (62–72 dph; shared neurons,  $n = 22$ ; specific neurons,  $n = 83$ ). Note that shared neurons had the same 'most-probable IBI' as neurons recorded during the protosyllable stage. Neurons exhibiting an increased burst period by skipping cycles of an underlying rhythm were also observed in birds 3, 4 and 6 (see Extended Data Figs 8f–h and 9f, h).





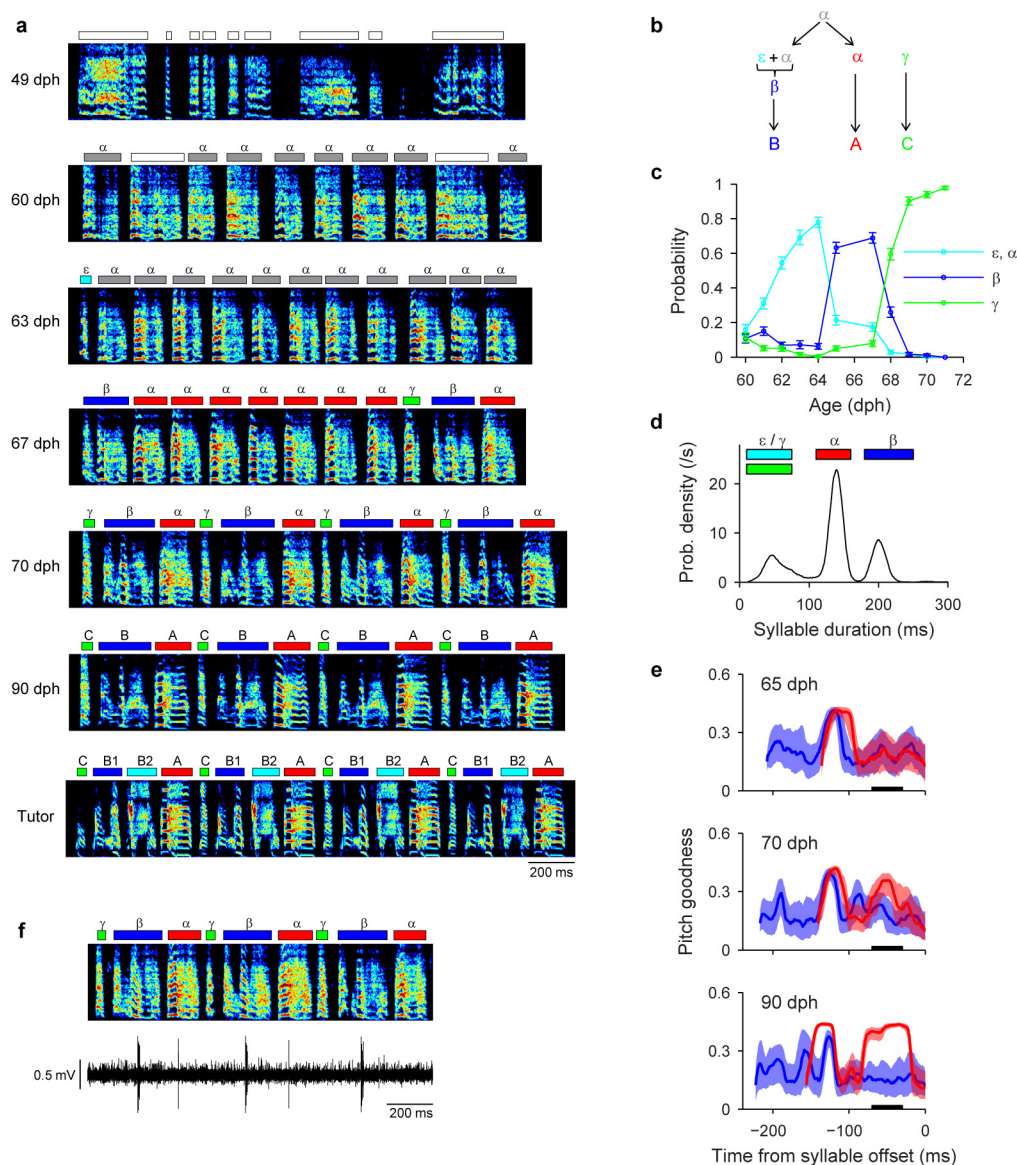
**Extended Data Figure 4 | Analysis of shared neurons: latency and syllable selectivity.** **a–d**, Latencies of shared neuron bursts, colour-coded by cell type:  $\text{HVC}_{\text{RA}}$  (red square),  $\text{HVC}_{\text{X}}$  (blue circle), and  $\text{HVC}_{\text{P}}$  (green diamond). **a**, Neurons in bird 1 shared between syllables  $\beta$  and  $\gamma$  (from Fig. 3) recorded during the early (top) and late (bottom) stages of syllable differentiation. Note strong correlation of burst latencies (early,  $r = 0.91$ ,  $P < 0.001$ ; late,  $r = 0.87$ ,  $P = 0.005$ ). **b**, Neurons in bird 1 shared between syllables D and B (Extended Data Fig. 7) during the early and late stages of syllable differentiation (top, early  $r > 0.99$ ,  $P < 0.001$ ; bottom, late  $r > 0.99$ ,  $P < 0.001$ ). **c**, Neurons in bird 2 shared between syllables  $\beta$  and  $\alpha$  (Fig. 4h) during the early and late stages (top, early  $r > 0.99$ ,  $P < 0.001$ ; bottom, late  $r > 0.99$ ,  $P < 0.001$ ). A shared neuron that had two peaks during the syllable  $\alpha$  is shown with an 'x' symbol; this point was not included in the calculation of correlation. **d**, Neurons in bird 4 shared between 'b<sub>2</sub>' and 'd<sub>1</sub>' (Extended Data Fig. 9l) during early stage (top,  $r = 0.89$ ,  $P < 0.001$ ; neurons that burst in the first part of 'b' ('b<sub>1</sub>') are shown with 'x' symbol, and were not included in the calculation of correlation). Neurons in bird 4 shared between syllables 'c' and 'd<sub>2</sub>' (Extended Data Fig. 9n) during early stage (bottom,  $r = 0.98$ ,  $P < 0.001$ ). Regarding bias: as a population, shared neurons exhibited a broad range of selectivity for emerging syllable types—some were equally active for both syllable types while others showed higher activity in one syllable than the other ('bias'; see Methods). **e**, Raw spike data (top left) and instantaneous firing rate (bottom left) for a neuron shared between syllables

$\beta$  and  $\gamma$  ( $\text{HVC}_{\text{P}}$ ; 68 dph, bird 1). Also shown is the syllable-onset-aligned raster plot (bottom right) and histogram (top right) showing similar peak firing rates for both syllables (low bias; bias = 0.07). **f**, Spike data (left) and syllable-onset-aligned raster plot and histogram (right) for a high-bias shared neuron showing higher peak firing rate for syllable  $\beta$  than  $\gamma$  (bias = 0.63;  $\text{HVC}_{\text{RA}}$ ; 68 dph, bird 1). **g**, Low-bias shared neuron (bias = 0.06;  $\text{HVC}_{\text{X}}$ ; 69 dph, bird 2). **h**, High-bias shared neuron showing higher peak firing rate for syllable  $\beta$  than  $\alpha$  (bias = 0.55;  $\text{HVC}_{\text{X}}$ ; 68 dph, bird 2). **i**, Scatter plot of the peak firing rates during two different syllable types, quantified by the height of the peak in the syllable-aligned spike histogram. Each dot is a neuron; shared neurons shown in cyan; neurons near the diagonal have low bias. Specific neurons are coloured according to the associated syllable and appear near the axes. **j**, Distribution of the bias for shared neurons (cyan) and specific neurons (magenta). Bias ranged from 0, representing equal activity, to 1, representing activity exclusive to either one of the syllables (see Methods). Specific neurons exhibited a bias tightly clustered around one ( $0.96 \pm 0.011$ , mean  $\pm$  s.d.). In contrast, shared neurons exhibited a broad range of bias ( $0.28 \pm 0.22$ ). These observations suggest that individual shared neurons can exist in a state intermediate between 'specific' and 'shared'—perhaps reflecting a gradual process by which shared neurons become specific. Scale bars for panels **e–h**, 0.5 mV, 100 ms. Insets in panels **f** and **h** show zoom of bursts indicated by an asterisk; scale bar: 5 ms.



**Extended Data Figure 5 | Analysis of the acoustic differences associated with shared neuron bursts.** While emerging syllable types gradually differentiate acoustically, some parts of different emerging syllable types may be acoustically quite similar. We wondered if shared neurons are only active at these times within emerging syllables at which no acoustic differentiation has yet occurred—that is, at times when the emerging syllable types are acoustically identical. To test this possibility, we analysed the trajectories of acoustic features of emerging syllable types around the times of shared neuron bursts. **a**, Shared HVC<sub>RA</sub> neuron recorded in bird 1 during alternation between emerging syllable types  $\beta$  and  $\gamma$  (same neuron as Fig. 3e). **b, c**, Average spectrogram (sparse contour representation; see Methods) computed for syllables  $\beta$  and  $\gamma$ , centred on a 50 ms window immediately after the burst in each syllable. **d**, Song amplitude as a function of time for syllables  $\beta$  (red) and  $\gamma$  (blue), relative to burst time. Lines show average across all syllable renditions on which the neuron was active. Shading around lines shows s.e.m. (for this and several other examples, s.e.m. is too small to be visible). **e**, Spectral centre of gravity as a function of time for syllables  $\beta$  (red) and  $\gamma$  (blue). **f**, Distribution of projected samples for syllables  $\beta$  (red) and  $\gamma$  (blue), computed by projecting the 8-dimensional vector of spectral features onto a line that yields maximum separability between the two syllables. This distribution is computed at each time (1 ms steps) in the 50-ms analysis window after burst time. Shown is the distribution at  $t = 25$  ms. **g**,  $d'$ -prime analysis of separability of projected samples for syllables  $\beta$  and  $\gamma$ . The value of  $d'$  is computed as a function of time (1 ms steps; red trace). Also shown is the 95% confidence interval (grey band) computed from surrogate data sets with randomized labels. Dashed horizontal line shows the 95 percentile of the distribution of peak

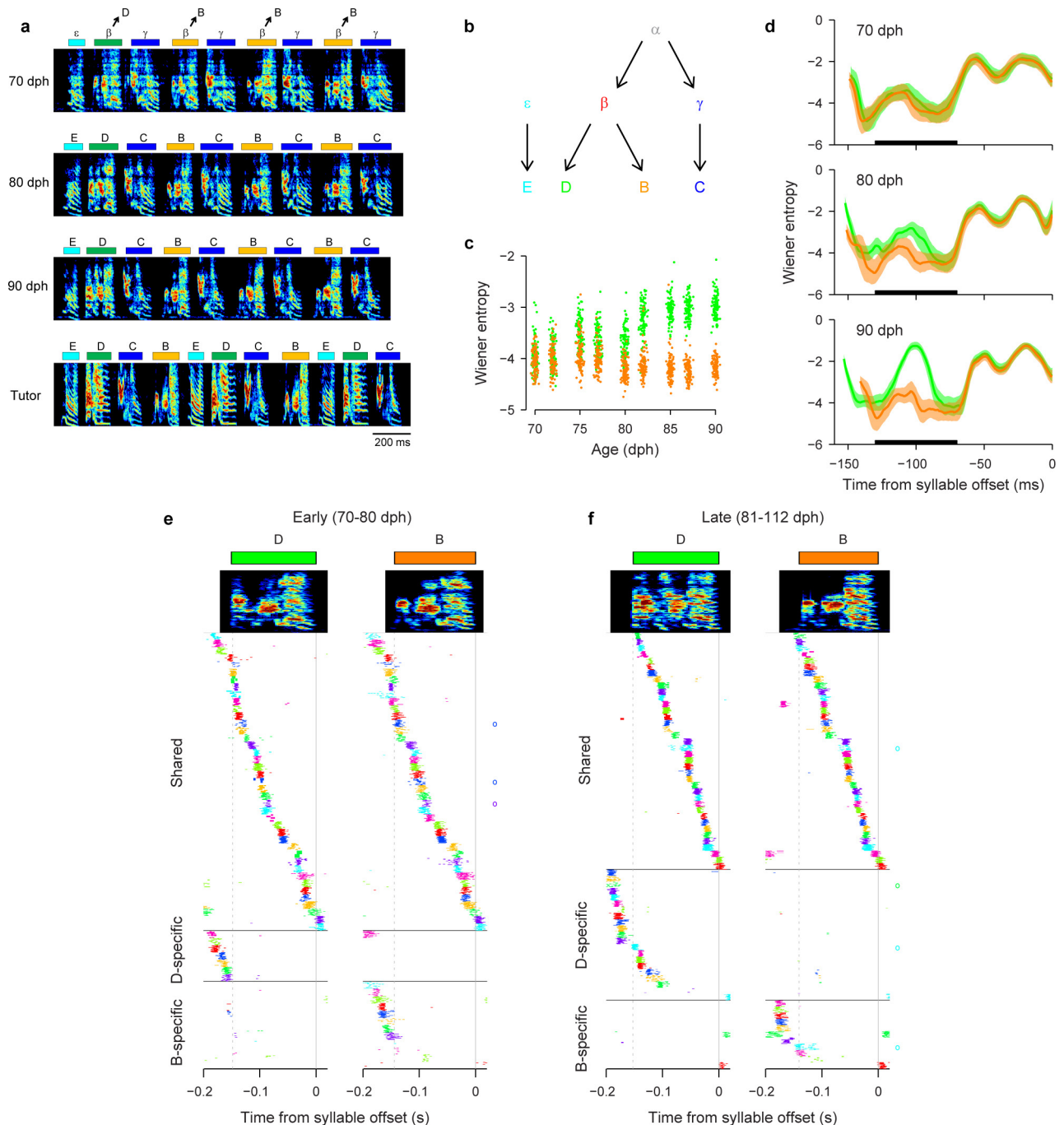
values of  $d'$  in the surrogate data set (identified in the 10–40 ms window). **h–j**, Acoustic analysis for three additional HVC<sub>RA</sub> neurons (analogous to panels a–g). **k**, Plot of  $d'$  trajectories for all shared HVC<sub>RA</sub> neurons. Significant  $d'$  values (above the 95 percentile of peak values) are shown in red. Non-significant values shown as grey lines. **l**, Same as panel k but for shared HVC<sub>X</sub> neurons. **m**, Population summary of mean  $d'$  (averaged over the presumptive premotor window 10–40 ms after burst time). Each symbol represents a different shared neuron and each column indicates a different syllable pair. Analysis is shown separately for each neuron type: HVC<sub>RA</sub> neurons (green circles) and HVC<sub>X</sub> neurons (blue squares). Neurons with no significant acoustic differences are indicated with black symbols. **n**, Cumulative distribution of mean  $d'$  for shared HVC<sub>RA</sub> neurons (green;  $n = 11$ ) and shared HVC<sub>X</sub> neurons (blue;  $n = 36$ ). Only neurons with significant  $d'$  metric are included in the cumulative. No significant difference was observed between neuron types ( $P = 0.1$ ). Scale bars for panels a, h, i, j are 0.5 mV, 100 ms. Summary of properties of HVC<sub>RA</sub> and HVC<sub>X</sub> shared neurons: Shared neurons were found in similar proportion across both HVC<sub>RA</sub> and HVC<sub>X</sub> neurons (19% and 28%, respectively;  $P = 0.08$ ; averaged over all developmental stages) and shared neurons of both cell types exhibited the property that bursts have similar latencies during the shared syllables (Extended Data Fig. 4a–d). As shown above, for both neuron types, we observed shared neurons that burst at times where there was a significant acoustic difference between the shared syllables. These findings suggest that both projection neuron types participate in shared neural sequences, and that these shared sequences occur during acoustically distinguishable parts of the emerging syllables.



**Extended Data Figure 6 | Detailed analysis of bout-onset differentiation in bird 2.** (Same bird as in Fig. 4). **a**, Song examples throughout song development. Panels from top to bottom: first, subsong (49 dph); second, emergence of protosyllable  $\alpha$  from subsong (60 dph); third, appearance of bout-onset element  $\epsilon$  (63 dph); fourth, fusion of  $\epsilon$  with first  $\alpha$  to form new syllable  $\beta$  (67 dph); fifth and sixth, acoustic differentiation of  $\beta$  and  $\alpha$ , and incorporation with  $\gamma$  into song motif CBA (70, 90 dph); seventh, tutor song. **b**, Schematic of syllable formation (same as Fig. 4a), inferred by tracking backward in development the adult syllables C, B and A. Early on, protosyllable (labelled  $\alpha$ ) is produced rhythmically. The first protosyllable in each bout fuses with a brief bout-onset vocal element  $\epsilon$  to form a new emerging syllable type  $\beta$ . Both  $\alpha$  and  $\beta$  undergo subsequent acoustic differentiation to form adult syllables A and B, respectively.

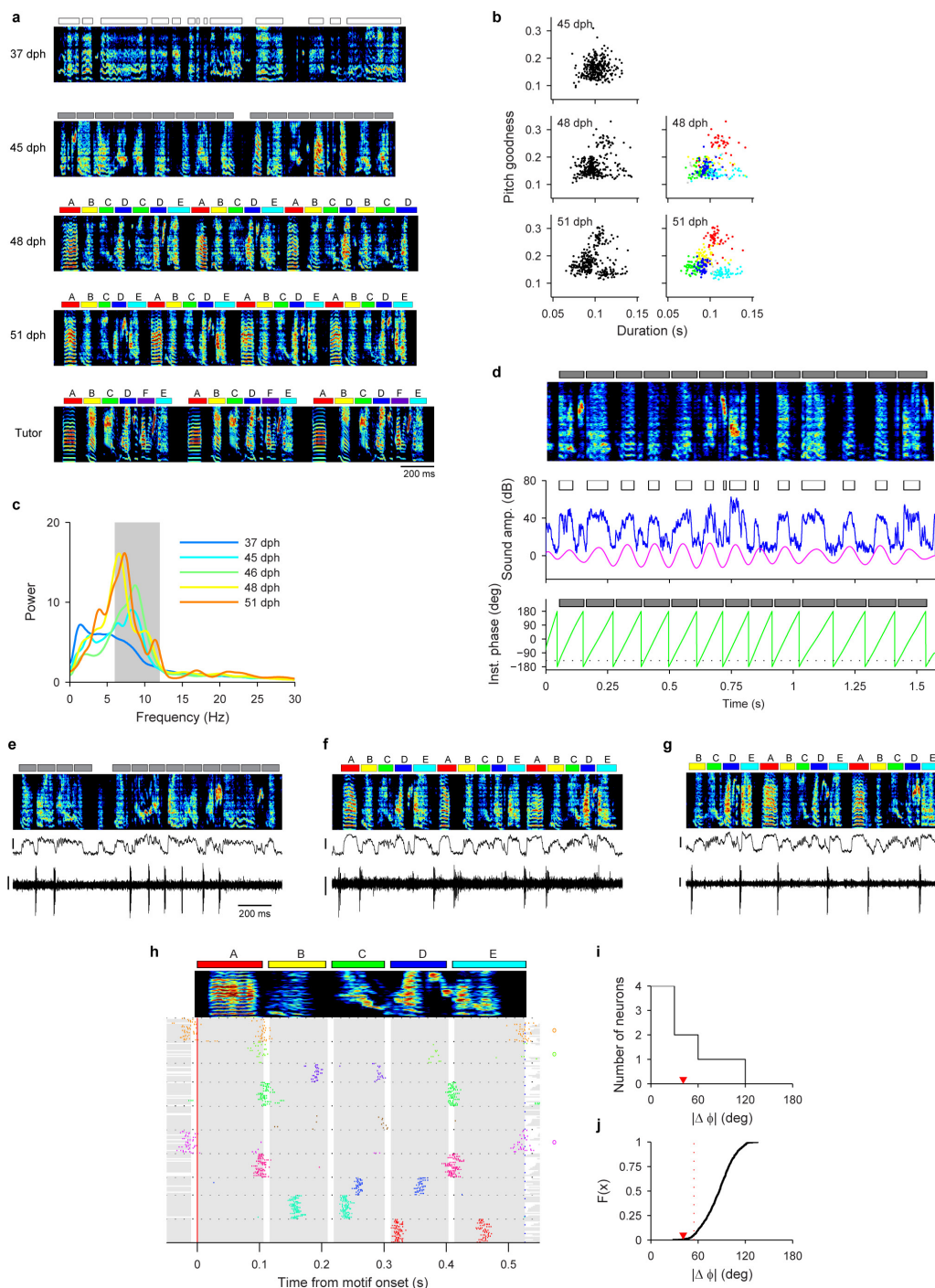
(An additional syllable  $\gamma$  emerges at bout onset to form adult syllable C). **c**, Developmental time course of the occurrence probability of different syllable types at bout onsets (mean  $\pm$  s.e.m.). **d**, Syllable duration distribution showing three non-overlapping peaks (67 dph). Coloured bars indicated syllable duration ranges used for syllable labelling. This separation of durations allowed automatic determination of syllable identity. **e**, Pitch goodness trajectories of syllables  $\alpha$  (red) and  $\beta$  (blue) at three stages of vocal development (median  $\pm$  quartiles;  $n = 100$  syllables per day). Black bar, region used to compute data in Fig. 4b. **f**, Example of a neuron active during both syllables  $\alpha$  and  $\beta$  (HVC<sub>RA</sub>; 69 dph). Note that the activity of this neuron during syllable  $\alpha$  was weak, and did not quite reach our statistical criterion for being a 'shared' neuron.





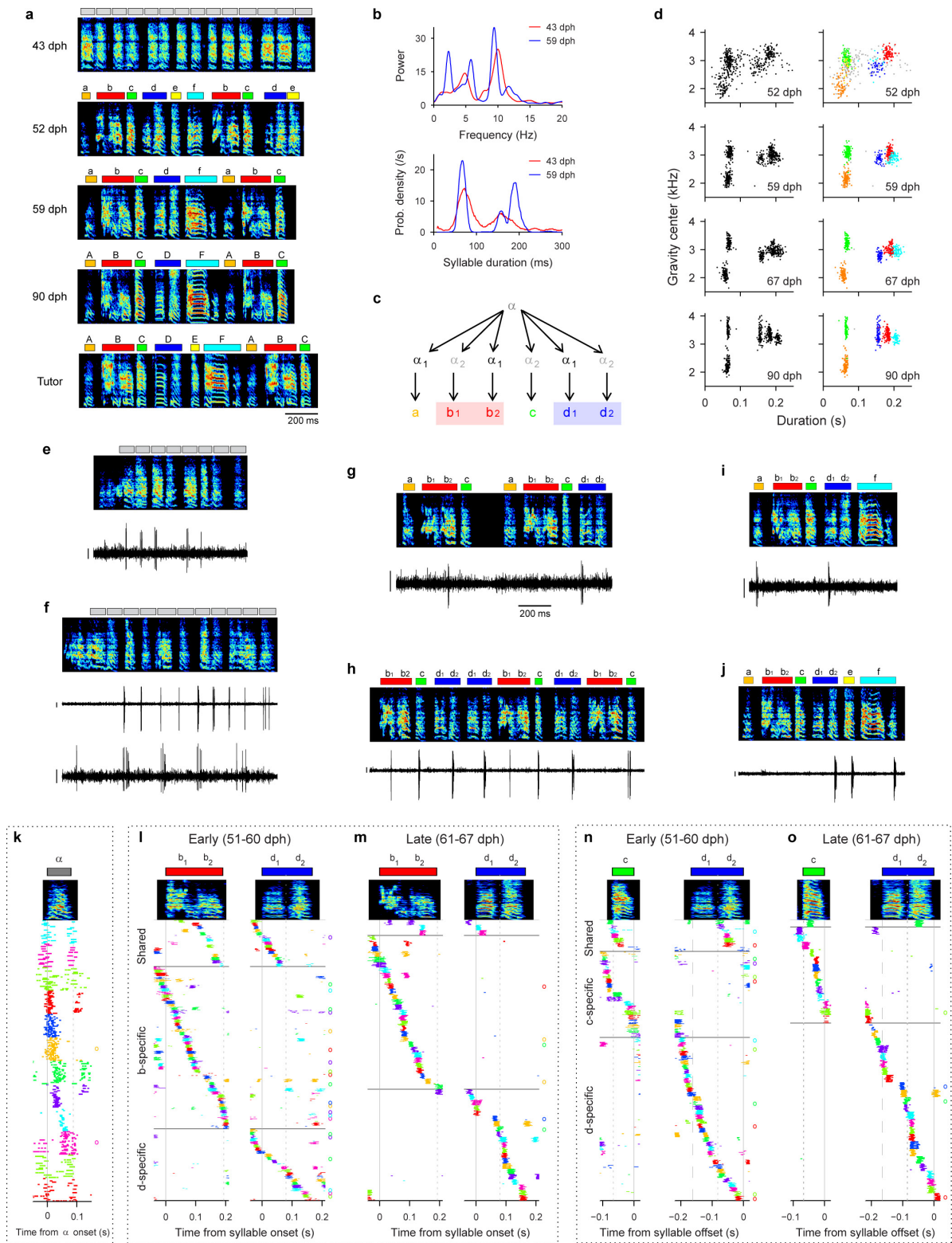
**Extended Data Figure 7 | Hierarchical differentiation of syllables.** All data are from bird 1 (same bird as in Fig. 3). **a**, Song examples during the emergence of syllables B and D from a common precursor syllable  $\alpha$ , which had undergone earlier differentiation from a protosyllable  $\beta$ . Panels from top to bottom: first (70 dph), After the initial differentiation of the protosyllable into  $\beta$  and  $\gamma$  (at  $\sim 62$  dph), the bird produced a rhythmic alternation of these two syllables, and the alternating sequence was reliably preceded at bout onsets by a short vocal element  $\epsilon$  ( $\epsilon$ - $\beta$ - $\gamma$ - $\beta$ - $\gamma$ - $\beta$ - $\gamma$ ...). Note that the first repetition of  $\beta$  in each bout (labelled D) is acoustically identical to later repetitions (labelled B); second (80 dph), the first repetition of  $\beta$  in the bout (syllable D) undergoes differential acoustic refinement compared to later repetitions (syllable B); third, syllable B, C and D, together with bout-onset element  $\epsilon$ , crystallize into adult motif EDCB (90 dph), that approximately matches the tutor motif (bottom panel). **b**, Schematic of syllable formation. **c**, Scatter plot of the mean Wiener entropy showing differential acoustic refinement of syllables B (orange) and D (green) through development ( $n = 100$  syllables of each type per day; horizontal jitter added to improve data visibility). **d**, Wiener entropy trajectory of syllables B and D at three stages of vocal development (median  $\pm$  quartiles;

$n = 100$  syllables of each type per day). Black bar indicates region used to compute data in panel **c**. **e**, Population raster of 60 neurons early in syllable differentiation showing shared (top) and specific (bottom) sequences. **f**, Same as **e**, but for 70 neurons recorded late in differentiation of D and B. Evidence for an incomplete splitting of a neural sequence: the pattern of shared and specific neurons observed for these syllables is quite similar to what would be expected in our model during an early/intermediate stage of splitting (Fig. 5c or Extended Data Fig. 10c). Of particular note in this bird is the large fraction of shared neurons between B and D that remained in the later recordings (panel **f**), compared to the smaller fraction of shared neurons at late stages in syllables B and C of the same bird (Fig. 3h). However, syllables B and C differentiated from parent syllable  $\alpha$  early in development ( $\sim 60$  dph, Fig. 3b), while D and B differentiated from  $\beta$  at a much later stage ( $\sim 80$  dph, panel **c**). One might speculate that the splitting of D and B may have failed to reach completion before the bird reached adulthood, possibly preventing further splitting. Neural evidence (shared burst sequence) for hierarchical differentiation was also observed in bird 6 (data not shown). Neural evidence (shared burst sequence) for bout-onset differentiation was also observed in bird 5 (data not shown).



**Extended Data Figure 8 | Simultaneous formation of multiple syllable types into an entire motif.** All data are from bird 3. Neural recordings from this bird support the view that, in the ‘motif strategy’, new syllables emerge from a common rhythmic protosequence. **a**, Song examples during the emergence of a motif. Panels from top to bottom: first, subsong (37 dph); second, the song began to acquire rhythmic ‘protosyllable’ modulation in song amplitude around 9 Hz (45 dph); third, over the next five days (47–51 dph), this bird acquired a reliable pattern of 4–5 acoustically distinct elements (‘syllables’), each generated in a different cycle of the 9 Hz rhythm (48 dph); fourth, the acoustic structure in each syllable was gradually refined, resulting in an excellent match to the tutor song even at this early age (51 dph); fifth, tutor song. **b**, Scatter plot of syllable duration and pitch goodness ( $n = 300$  syllables per day; colour coded according to syllable identity in panel **a**). **c**, Development of song rhythmicity quantified as the spectrum of the sound amplitude<sup>38</sup>. Gray shade indicates the pass band for the filter used in phase segmentation. **d**, Phase segmentation based on the rhythmicity in the song. Top, song spectrogram with phase segments (grey boxes). Middle, sound amplitude (blue) and band-pass filtered sound

amplitude (magenta). Syllable segmentation based on the sound amplitude is shown as white boxes. Bottom, instantaneous phase (green) of the band-pass filtered sound amplitude. Phase segments (grey boxes) are obtained by detecting threshold crossing (black dotted line) of the instantaneous phase. **e**, Rhythmic neuron (protosyllable stage; HVC<sub>P</sub>; 45 dph). **f**, Neuron shared between syllables A and B (HVC<sub>RA</sub>; 48 dph). **g**, Neuron shared between B and E (HVC<sub>X</sub>; 49 dph). **h**, Population raster aligned to the five-syllable motif for neurons that were significantly locked to any syllable ( $n = 10$  neurons). Each motif and associated spike times were time-warped using a piecewise linear method<sup>67</sup> based on syllable onsets and offsets. **i**, Histogram of the absolute phase difference between the two syllables for all shared neurons ( $n = 8$  neurons; mean phase difference:  $41 \pm 33.9$  deg, mean  $\pm$  s.d.). **j**, Cumulative distribution of the mean absolute phase difference after randomizing burst identity (red dotted line indicates  $P = 0.05$  threshold for significance; red triangle indicates observed mean absolute phase difference,  $P = 0.013$ ). Statistical details in Methods. Scale bars for panels **e–g**, 30 dB, 0.3 mV, 200 ms.



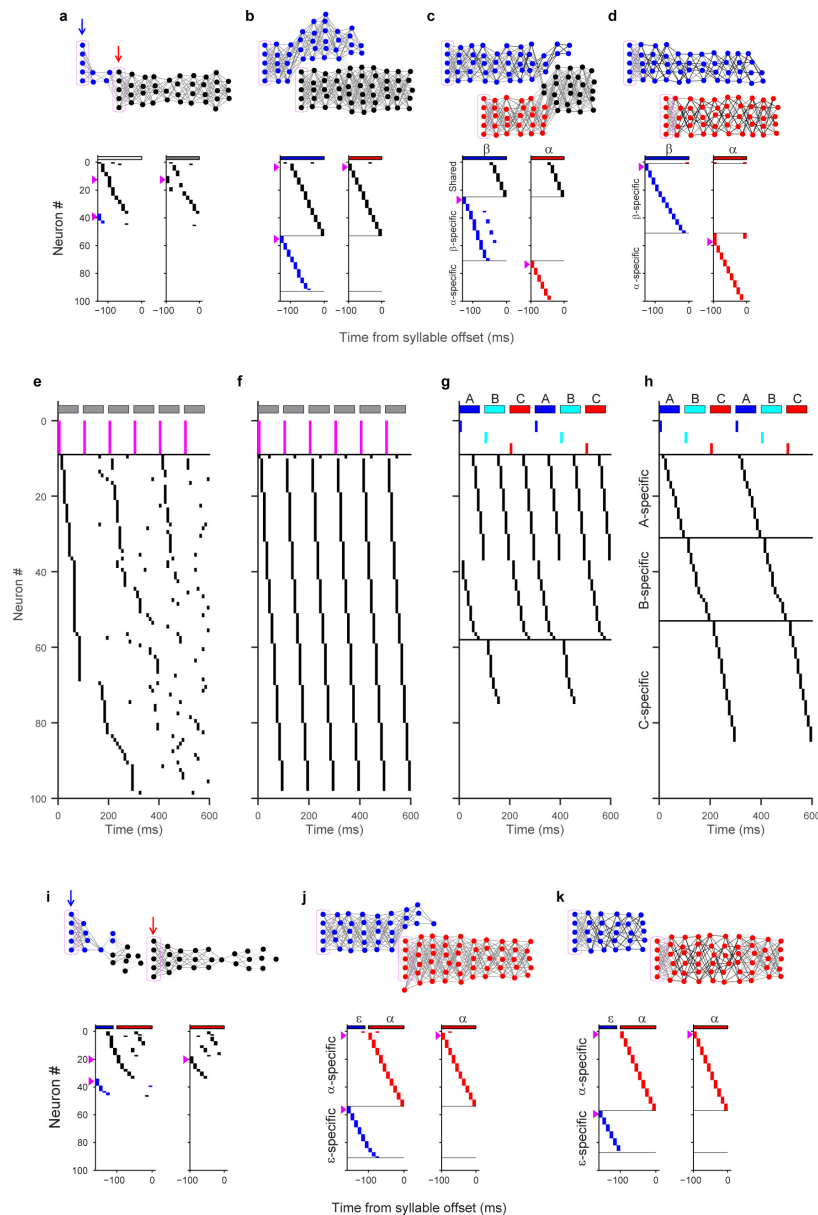
Extended Data Figure 9 | See next page for caption.



# Extended Data Figure 9 | Another example of shared burst sequences during the emergence of new syllable types. All data are from bird 4.

**a**, Song examples during the emergence of a motif ABCDE. Note the nearly simultaneous emergence of multiple syllable types in nearly fixed order (52 dph). Tutor song shown at the bottom. Phase segments are shown above the spectrogram for song at 43 dph. **b**, Top, song rhythm spectrum calculated in the protosyllable stage (43 dph) and after motif formation (59 dph). Note the pronounced peaks at 5 Hz and 10 Hz in both stages. Bottom, syllable duration distribution in the protosyllable stage (43 dph) and after motif formation (59 dph) showing two peaks. At 43 dph, the peak at 70 ms indicates short protosyllables corresponding to one cycle of the 10 Hz rhythm, and the peak at 140 ms indicates longer syllables formed by two protosyllables fused across two cycles of the 10 Hz rhythm (doubled protosyllables). Example doubled protosyllables are seen in the first and third syllables of panel **a**, 43 dph. (Note that boxes at the top of this panel indicate phase segments, not syllable boundaries). **c**, Hypothesized mechanism of motif construction, based on the examination of acoustic structure and analysis of neural burst sequences (see below). Notably, in this bird, the majority of syllables emerged nearly simultaneously in a relatively fixed order, consistent with a 'motif strategy'. **d**, Scatter plots of syllable duration versus mean spectral centre of gravity at four stages of vocal development (each dot represents a single syllable;  $n = 500$  syllables per day; colour coded according to syllable identity in panel **a**). **e**, Neuron bursting at the 10 Hz protosyllable rhythm (HVC<sub>X</sub>; 48 dph). Phase segments shown above spectrogram. **f**, Top, neuron bursting at the 10 Hz rhythm (HVC<sub>X</sub>; 49 dph). Bottom, simultaneous recording of a neuron bursting on alternate cycles of the 10 Hz rhythm (HVC<sub>RA</sub>). **g**, Shared neuron bursting on second half of syllable 'b' (labelled  $b_2$ ) and first half of syllable 'd' (labelled  $d_1$ ) (HVC<sub>RA</sub>; 51 dph). **h**, Shared neuron bursting rhythmically on 'b<sub>1</sub>', 'c' and second half of 'd' ( $d_2$ ) (HVC<sub>RA</sub>; 51 dph). **i**, Shared neuron bursting on 'a' and 'd<sub>1</sub>' (HVC<sub>RA</sub>; 58 dph). **j**, Shared neuron bursting on 'd<sub>2</sub>', 'e', and last part of 'f' (HVC<sub>RA</sub>; 57 dph). **k**, Population raster of 12 neurons that were significantly locked to protosyllable onsets (48–49 dph). Protosyllables were identified using phase segmentation (see Methods). **l**, Population raster showing neurons active during syllables 'b' and/or 'd', recorded early in syllable differentiation. Neurons shared between 'b' and 'd<sub>1</sub>' are grouped at top. Neurons specific for 'b' are grouped next, and neurons specific for 'd' are grouped at bottom. **m**, Same as panel **l**, but for neurons recorded later in development. **n**, Population rasters showing neurons active during syllables 'c' and/or 'd', recorded early in development. **o**, Same as **m**, but for neurons recorded later in development. Scale bars for panels **e–j**, 0.5 mV, 200 ms. Neural evidence for hypothesized mechanism of motif construction: based

on an analysis of acoustic signals and neural recordings, we have formulated a hypothesis for how the song of this bird developed, from the formation of the protosyllable to the emergence of the complete motif. We hypothesize that the fundamental protosyllable element corresponds to the prominent 10 Hz peak in the rhythm spectrum and the 70 ms peak in the duration distribution (panel **b**). This view is further supported by the presence of neurons in the protosyllable stage that generate rhythmic bursts at 10 Hz (panels **e** and **f**; 11/18 neurons were rhythmic, 5/11 rhythmic neurons exhibited periodicity at 10 Hz), and the existence of a burst sequence during the protosyllable (panel **k**). In this bird, the rhythmic protosyllables differentiated nearly simultaneously, at an early age (52 dph, panel **a**), into a complete sequence of distinct syllables that subsequently formed the adult song, suggesting this bird employed a 'motif strategy'. One complication of this simple view is that there may have been an early partial splitting of the short protosyllable  $\alpha$  into two 'daughter' protosyllables  $\alpha_1$  and  $\alpha_2$ , which alternated to produce the elements of the final motif (panel **c**). Two lines of evidence based on neural activity support this view: First, many neurons recorded at an early stage ( $< 50$  dph) exhibited a prominent 5 Hz periodicity in their rhythmic bursting, (panels **f** and **h**; 6/11 rhythmic neurons), rather than the expected 10 Hz period (panels **e** and **f**, top trace). This observation led us to consider the possibility that the 100 ms neural sequence, corresponding to the dominant 10 Hz protosyllable rhythm, underwent a partial splitting during the protosyllable stage—similar to the alternating differentiation described for bird 1 (Fig. 3; Extended Data Fig. 4). This would result in two distinct alternating protosyllable sequences  $\alpha_1$  and  $\alpha_2$  (panel **c**). Such splitting would effectively double the period of the protosyllable rhythm, and would account for the 'doubled' protosyllables and the 5 Hz peak in the rhythm spectrum (panel **b**). The existence of short and doubled protosyllables led us to hypothesize that the short syllables of the adult motif ('a', 'c', and 'e') arose from the short protosyllables, while long adult syllables ('b' and 'd', and possibly 'f') arose from the doubled protosyllables (panel **c**). Early syllable 'e' is later dropped by the juvenile, although it appears in the tutor song. Furthermore, the analysis of shared sequences (panels **l–o**) revealed a predominance of shared neurons between syllable elements in alternating cycles of the underlying 10 Hz rhythm. For example, shared neurons were observed between syllables 'a', 'b<sub>2</sub>' and 'd<sub>1</sub>' (panel **i** for neuron shared between 'a' and 'd<sub>1</sub>'; panels **g** and **l** for neurons shared between 'b<sub>2</sub>' and 'd<sub>1</sub>'). Shared neurons were also observed between syllables 'b<sub>1</sub>', 'c', and 'd<sub>2</sub>' (panel **h** for neuron shared between 'b<sub>1</sub>', 'c', and 'd<sub>2</sub>'; panel **n** for neurons shared between 'c' and 'd<sub>2</sub>'). In contrast, many fewer shared neurons were observed between neighbouring cycles of the underlying rhythm, although examples of this can be found (panel **j**).



**Extended Data Figure 10 | Model of other strategies for syllable formation.** **a–d**, Bout-onset differentiation results from activation of bout-onset seed neurons (blue arrow) followed by rhythmic activation of protosyllable seed neurons (red arrow). Network diagrams show **(a, b)** protosyllable formation and **(c, d)** splitting of chains specific for bout-onset syllable  $\beta$  and specific for later repetitions of the protosyllable  $\alpha$  (blue and red, respectively; shared neurons: black). **e–h**, Model of simultaneous formation of multiple syllable types into an entire motif ('motif strategy'). **e, f**, Protosyllable seed neurons (magenta lines) were activated rhythmically to form a protosequence. **g**, Seed neurons were then divided into three sequentially activated subgroups, resulting in the rapid splitting of the protosequence into three daughter sequences. In intermediate stages (panel **g**), individual neurons exhibited varying degrees of specificity and sharedness for the emerging syllable types. **h**, After learning, the population of neurons was active sequentially throughout the entire 'motif', but individual neurons were active during only one of the resulting syllables, forming three distinct non-overlapping sequences. **i–k**, Network diagrams and raster plots showing an example of the formation of a new syllable chain at bout onset. In the network diagrams, seed neurons are indicated within magenta boxes, and bout-onset seed neurons and protosyllable seed neurons are indicated by blue and red arrows, respectively. Neurons specific for each emerging syllable type ( $\varepsilon$  and  $\alpha$ ) are coloured blue and red, respectively. The three panels represent the early protosyllable stage, the late protosyllable stage, and the final stage. The training protocol is similar to that for bout-onset differentiation (panels **a–d**), except that protosyllable seed neurons

are driven more strongly throughout the learning process. As a result, protosyllable seed neurons did not become outcompeted by the growing bout-onset chain. Strong activation of the protosyllable seed neurons also terminated activity in the bout-onset chain through fast recurrent inhibition, thus preventing further growth of the bout-onset chain, as occurs in bout-onset differentiation. Regarding the role of chain splitting in the formation of new syllable types: in our model, we envision that the formation of daughter chains in HVC is translated into the emergence of new syllable types as follows. During the splitting process, as two distinct sequences of specific neurons develop, their downstream projections can be independently modified<sup>67,77</sup> such that each of the emerging chains of specific neurons can drive a distinct pattern of downstream motor commands, allowing distinct acoustic structure in the emerging syllable types. Such differential acoustic refinement is consistent with the previous behavioural observation that the altered acoustic structure of new syllables emerges in place, without moving or reordering sound components ('sound differentiation *in situ*')<sup>33</sup>. This model naturally explains the apparent 'decoupling' of shared projection neuron bursts from acoustic structure in the vocal output—that is, the fact that the bursts of shared neurons become associated with two distinct acoustic outputs during the differentiation of two syllable types (Extended Data Fig. 5). Specifically, during syllable differentiation, a shared neuron participates with different ensembles of neurons during each of the emerging sequences, and these different ensembles can drive different vocal outputs.

# Acute off-target effects of neural circuit manipulations

Timothy M. Otchy<sup>1,2</sup>, Steffen B. E. Wolff<sup>1\*</sup>, Juliana Y. Rhee<sup>1\*</sup>, Cengiz Pehlevan<sup>3\*</sup>, Risa Kawai<sup>1</sup>, Alexandre Kempf<sup>1†</sup>, Sharon M. H. Gobes<sup>1†</sup> & Bence P. Ölveczky<sup>1,4</sup>

**Rapid and reversible manipulations of neural activity in behaving animals are transforming our understanding of brain function. An important assumption underlying much of this work is that evoked behavioural changes reflect the function of the manipulated circuits. We show that this assumption is problematic because it disregards indirect effects on the independent functions of downstream circuits. Transient inactivations of motor cortex in rats and nucleus interface (Nif) in songbirds severely degraded task-specific movement patterns and courtship songs, respectively, which are learned skills that recover spontaneously after permanent lesions of the same areas. We resolve this discrepancy in songbirds, showing that Nif silencing acutely affects the function of HVC, a downstream song control nucleus. Paralleling song recovery, the off-target effects resolved within days of Nif lesions, a recovery consistent with homeostatic regulation of neural activity in HVC. These results have implications for interpreting transient circuit manipulations and for understanding recovery after brain lesions.**

Understanding how the brain generates behaviour is a daunting task often simplified by studying anatomically distinct brain regions in isolation. The underlying assumption is that different parts of the brain are specialized for different functions that can be understood by monitoring and altering activity in local circuits. An increasingly powerful and widely used approach is to transiently silence or otherwise perturb—by optogenetic, pharmacological or other means—neural activity in specific circuits and observe the consequences on behaviour<sup>1,2</sup>. If there is an effect, the conclusion is that the circuit under investigation is causally ‘involved’ in the behaviour. But what does such a causal link actually tell us?

In a densely interconnected dynamical system like the brain, sudden perturbations to one node (for example, a brain area) could send ripples through the system, compromising the capacity of downstream circuits to perform computations on other inputs or generate patterned activity from internal dynamics<sup>3,4</sup>. Given the reliance on transient circuit manipulations for localizing computations and memory functions to specific neural circuits or brain areas<sup>1,2</sup>, the caveats and limitations of these methods should be scrutinized.

If inactivating a brain area interferes with the independent functions of downstream circuits, that is, functions not contingent on information provided by the targeted area, an important next question is whether those functions remain compromised after the silencing is made permanent through lesions. For example, deficits caused by changes in the excitability of downstream neurons could plausibly resolve through homeostatic regulation of neural activity<sup>5–9</sup>. Spontaneous recalibration of neural dynamics, more generally, could help explain why chronic effects of permanent lesions are often far less severe than those induced by transient inactivations<sup>10–12</sup>, and why patients with strokes and other brain injuries can overcome some of their initial deficits without rehabilitation<sup>13</sup>.

Functional recovery after brain lesions, however, is thought to be driven predominantly by the adoption of new behavioural strategies and the adaptive repurposing of non-lesioned circuits<sup>12,14</sup>, processes

contingent on renewed experience with affected tasks<sup>15</sup>. Therefore, demonstrating acute off-target effects of inactivations and spontaneous recovery after permanent lesions requires showing that task-specific behaviours sensitive to transient inactivations can recover after lesions without additional task experience. As experience-dependent recovery is difficult to rule out for basic sensory or motor functions that are central to many behaviours and hence naturally ‘practiced’ after lesions<sup>10,11,14</sup>, our study used behaviours for which such incidental practice can be withheld. We chose learned movement sequences of rats<sup>16</sup> and courtship songs of zebra finches<sup>17</sup> because they are task-specific skills associated with complex, stereotyped and idiosyncratic motor patterns that can be precisely quantified and compared across various manipulations.

To probe the effects of transient manipulations on distinct and independent functions of downstream circuits, we targeted brain areas—motor cortex in rats and the sensorimotor area Nif in songbirds—that are known, based on lesion studies, to be dispensable for storing and executing the skills we study<sup>16,18</sup>. Despite this, transient manipulations severely degraded the learned behaviours in both systems. These discrepancies were consistent with acute disruptions of downstream circuit function. Though we saw similar behavioural effects immediately after permanent lesions, they resolved spontaneously, leading to full recovery of the initially affected behaviours.

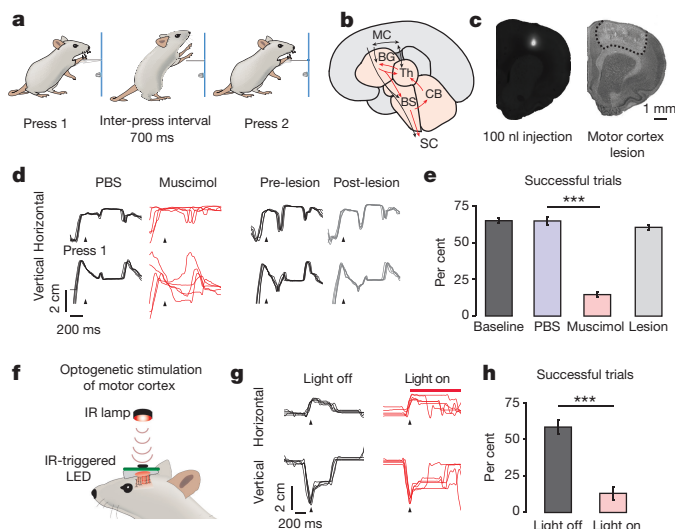
## Motor cortex inactivation disrupts skill execution

In our motor learning task, rats are rewarded for pressing a lever in a precise temporal sequence (two presses 700 ms apart, Fig. 1a). Animals solve this task by acquiring spatiotemporally precise movement patterns that produce the prescribed lever-press sequence<sup>16</sup>. Though the learned skills are robust to motor cortical lesions<sup>16</sup>, motor cortex projects to sub-cortical motor structures whose distinct functions could be sensitive to sudden changes in motor cortical input (Fig. 1b). To probe this, we inactivated primary forelimb motor cortex of rats that had learned the task ( $n = 5$  rats) by injecting 100 nl of the GABA<sub>A</sub> agonist

<sup>1</sup>Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA. <sup>2</sup>Program in Neuroscience, Harvard University, Cambridge, Massachusetts 02138, USA. <sup>3</sup>Center for Computational Biology, Simons Foundation, New York, New York 10010, USA. <sup>4</sup>Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. <sup>†</sup>Present addresses: Institut de Biologie, Ecole Normale Supérieure, Paris, France (A.K.); Neuroscience Program, Wellesley College, Wellesley, Massachusetts 02481, USA (S.M.H.G.).

\*These authors contributed equally to this work.





**Figure 1 | Motor skills that survive motor cortical lesions are acutely affected by transient manipulations of motor cortex.** **a**, We used a motor skill learning paradigm that trains rats to press a lever twice with a specified inter-press interval (IPI), typically 700 ms. **b**, Schematic of the mammalian motor system. Motor cortex (MC, black) provides input to subcortical circuits (red shaded regions, BG, basal ganglia; CB, cerebellum; BS, brainstem; Th, thalamus; SC, spinal cord). **c**, Coronal sections comparing the spread of a fluorescent marker (fluorescein) matched in concentration and volume to our muscimol injections (left) with motor cortex lesions that leave the learned skills intact<sup>16</sup> (right, dashed lines denote the lesioned area) (Methods). **d**, Left, forepaw trajectories associated with five consecutive trials after PBS and muscimol injections for the rat that received the lowest dose of muscimol (100 nl, 1 mM) (Supplementary Video 1). Right, motor cortex was subsequently lesioned in this rat. Paw trajectories associated with five consecutive trials from the last training session before and the first training session after the lesion. **e**, Fraction of trials with IPIs within 20% of the target ('successful' trials) for different experimental conditions ( $n = 5$  rats). Lesion data from ref. 16 shown for comparison (light-grey bar). **f**, Wireless optogenetic stimulation (Methods). **g**, Paw trajectories for five consecutive trials for an example rat with and without optogenetic stimulation of motor cortex. **h**, Same as **e**, but with and without optogenetic stimulation of motor cortex ( $n = 5$  rats). Error bars represent standard error of the mean (s.e.m.). \*\*\* $P < 0.001$ . For tests of significance for this and all other figures see Methods.

muscimol (1–25 mM) into the hemisphere contralateral to the dominant paw<sup>19,20</sup> (Methods). Based on previous studies<sup>21</sup> and injections of 100 nl of fluorescein into motor cortex (Fig. 1c), we estimate that the direct effects of our injections were confined to a volume far smaller than our previous lesions<sup>16</sup> (Fig. 1c; Methods).

In contrast to animals tested for the first time 5–10 days after lesions<sup>16</sup>, muscimol-injected rats had severe deficits in skill execution with marked drops in performance and disrupted paw kinematics (Fig. 1d, e). These effects were evident even in the rat receiving the lowest concentration of muscimol (1 mM) (rat in Fig. 1d, Supplementary Video 1). We later lesioned motor cortex in this rat, and as previously reported (rat 'Kansas' in ref. 16), saw no effect on skill execution when the rat was tested again 10 days post-lesion (Fig. 1d).

To explore dose-dependence, we injected larger volumes of muscimol in two of the rats (200 nl and 400 nl respectively). We found task performance to be even more affected with lever-interactions restricted to a few single presses with no rewarded trials (Supplementary Video 1).

### Optogenetic stimulation of motor cortex

Transient stimulation of neural activity is an alternative method for disrupting ongoing circuit dynamics that is well-suited for interrogating processes associated with precise and reproducible neural dynamics<sup>22</sup>. As with transient inactivations, sudden activation can also plausibly

affect the dynamics and function of downstream circuits. To probe the effect of transient motor cortex stimulation on skill execution, we used optogenetics, a widely adopted method for manipulating neurons in temporally specific ways<sup>23</sup>.

We expressed the optogenetic activator Chrimson<sup>24</sup> in motor cortex ( $n = 5$  rats; Extended Data Fig. 1a, b; Methods), and stimulated the hemisphere contralateral to the dominant paw after animals had reached asymptotic performance on the task (Fig. 1f and Extended Data Fig. 1c). Neither brief (50 ms) nor sustained (1 s) optogenetic stimulation evoked visible motor responses during rest, suggesting that they were sub-threshold for movement initiation. However, both brief and sustained stimulation, triggered on the first lever-press in a trial, interfered with task performance and associated kinematics (Fig. 1g, h, Extended Data Fig. 2 and Supplementary Video 2). Thus, similar to transient inactivations, disrupting normal activity patterns in motor cortex by optogenetic stimulation compromises the animals' capacity to execute skills robust to permanent lesions.

### Effects of Nif lesions and inactivations differ

Our results suggested that behavioural effects of transient perturbations may overestimate the steady-state functions of targeted circuits. To examine whether this caveat should be considered more broadly, we similarly probed whether transient inactivations of sensorimotor nucleus Nif in zebra finches affect their courtship songs. Although the song survives Nif lesions<sup>18</sup>, Nif sends excitatory projections to HVC, an essential part of the song control circuit believed to generate the temporal pattern for learned vocalizations through intrinsic network dynamics<sup>25</sup> (Fig. 2a and Extended Data Fig. 3a).

We first confirmed the findings of previous lesion studies<sup>18</sup> by injecting 27–36 nl of *N*-methyl-DL-aspartic acid (NMA), an excitotoxin, bilaterally into Nif ( $n = 5$  birds) (Fig. 2a and Extended Data Fig. 3b). When Nif lesioned birds resumed singing two days after surgery, their songs were similar to pre-lesion (Fig. 2b, c), consistent with prior studies. Because birds did not sing within the first day of lesions, this result does not preclude short-term effects of Nif silencing<sup>18</sup>. To probe such acute effects, we injected 27 nl of muscimol (50 mM) bilaterally into Nif of awake head-restrained adult birds ( $n = 5$  birds; Fig. 2d, e, Extended Data Fig. 4a; Methods).

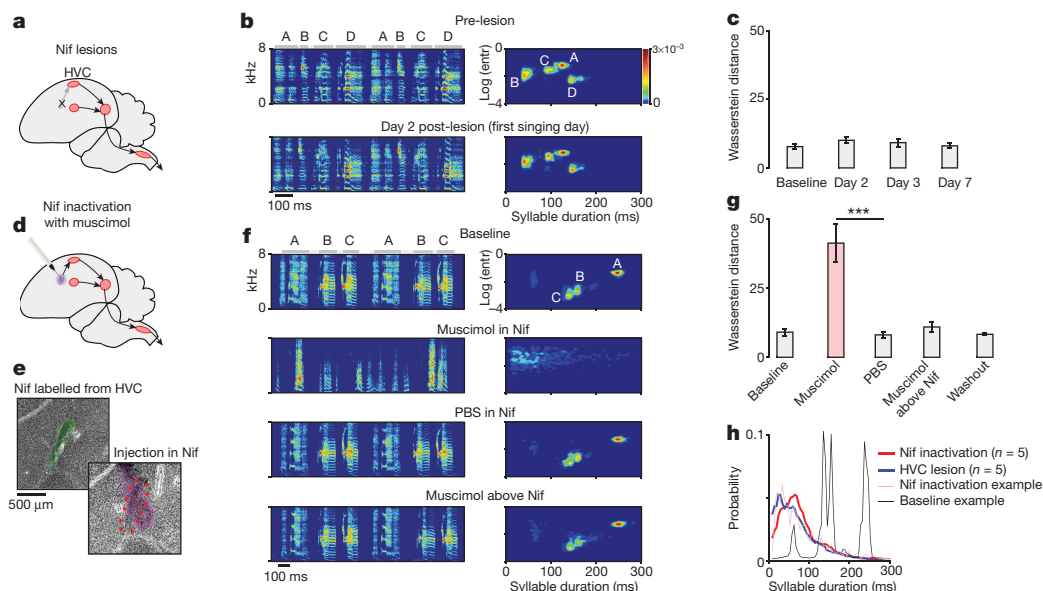
Birds typically sang within 20 min of the injections, but their songs were severely degraded and reminiscent of subsong (Fig. 2f, g), highly variable and unstructured utterances normally produced by juvenile birds at the start of vocal learning or by adult birds after bilateral HVC lesions<sup>26</sup>. The syllable duration distributions were similar to those reported in HVC-lesioned birds<sup>26</sup> (Fig. 2h), suggesting that Nif inactivation degrades song by indirectly affecting HVC dynamics.

To exclude the possibility that the behavioural effects were caused by diffusion of muscimol into HVC, we injected the same dose 300–500  $\mu$ m dorsal to Nif but closer to HVC ( $n = 5$  birds), as well as smaller volumes (9 nl) into Nif ( $n = 2$  birds). The control injections above Nif did not affect song (Fig. 2f, g), whereas the smaller Nif injections evoked effects similar to the larger dose (Extended Data Fig. 4b).

### HVC dynamics recover after Nif lesion

Transient circuit manipulations performed in both rats and songbirds revealed strong effects on skill execution not seen after permanent lesions (Figs 1 and 2). The discrepancy could not be explained by experience-dependent relearning in lesioned animals because the skills recovered their idiosyncratic pre-lesion form without any intervening practice (Figs 1d, e and 2b, c)<sup>16</sup>. However, the acute behavioural deficits were consistent with activity manipulations in motor cortex and Nif indirectly affecting the independent functions of downstream circuits. These initially affected circuits, however, seemingly regained their capacity to execute the learned behaviours after permanent lesions<sup>16,18</sup>.

Unlike in rats, where the neural circuits underlying the skills we assay have yet to be characterized, the vocal control circuits in zebra finches have been well delineated<sup>27</sup>, making it feasible to investigate



**Figure 2 | Transient inactivations of Nif severely degrade adult zebra finch song, while permanent lesions have no noticeable effect when singing resumes two days later.** **a**, Schematic of the song control circuit (red nuclei). Nif, a sensorimotor nucleus that inputs to HVC, was lesioned bilaterally ( $n = 5$  birds). **b**, Left, spectrograms show two song motifs (syllables ABCD) from an example bird before (top), and after (bottom) bilateral Nif lesion. Right, joint entropy-duration distributions for song syllables uttered before and two days after the lesion. Letters denote syllables in the bird's song motif. **c**, Summary statistics showing the difference (Wasserstein distance) between the joint entropy-duration distributions before and on different days after Nif lesions. 'Baseline' compares the distributions from two consecutive days of pre-lesion

downstream effects of local circuit manipulations. On the basis of our results and known anatomy (Fig. 2), we hypothesized that Nif inactivations perturb vocal output by removing excitatory input from HVC, thus compromising the function of this song-specialized premotor area<sup>28</sup>.

Most lesion protocols require surgery, which suppresses singing for a day or two, exactly the timeframe during which we hypothesize that recovery in HVC function occurs. To monitor neural dynamics in HVC in the immediate aftermath of Nif lesions and to compare it to pre-lesion dynamics, we lesioned Nif in freely behaving birds while recording multi-unit neural activity in HVC<sup>29</sup>. Stimulation electrodes targeted to Nif were implanted together with recording probes in ipsilateral HVC (Fig. 3a; Methods). Nif was lesioned unilaterally by injecting 50  $\mu$ A of current for 30–40 s ( $n = 11$  birds; Fig. 3b). Nif was successfully ablated (>80% lesioned; Methods) in 4 out of 11 birds, and subsequent analysis was done on this cohort unless otherwise noted.

Spontaneous (that is, non-vocal) HVC activity was dramatically reduced immediately following the lesions, consistent with a sudden loss of excitatory input from Nif<sup>30</sup>, but recovered in the ensuing hours (Fig. 3c). Singing, which was invariably interrupted by the electrical stimulation, resumed after  $1.3 \pm 0.9$  h (Fig. 3c–e). A fraction of the initial post-lesion vocalizations were severely degraded and did not resemble pre-lesion song<sup>18</sup> (Fig. 3d, e and Extended Data Fig. 5). The effects were less severe than during bilateral Nif inactivation (Fig. 2f, g), probably reflecting bilateral control of zebra finch song<sup>31</sup>.

For vocalizations that resembled pre-lesion song, neural activity was aligned to a common song template (Methods). Although song-aligned activity patterns were similar across renditions in the hours following Nif lesions, they were strikingly different from pre-lesion dynamics (Fig. 3f–h).

Despite the initial degradation of song and associated HVC dynamics, both gradually recovered (Fig. 3e–h). By the second day, the song was reliably back to pre-lesion form (Fig. 3e). Remarkably, the average

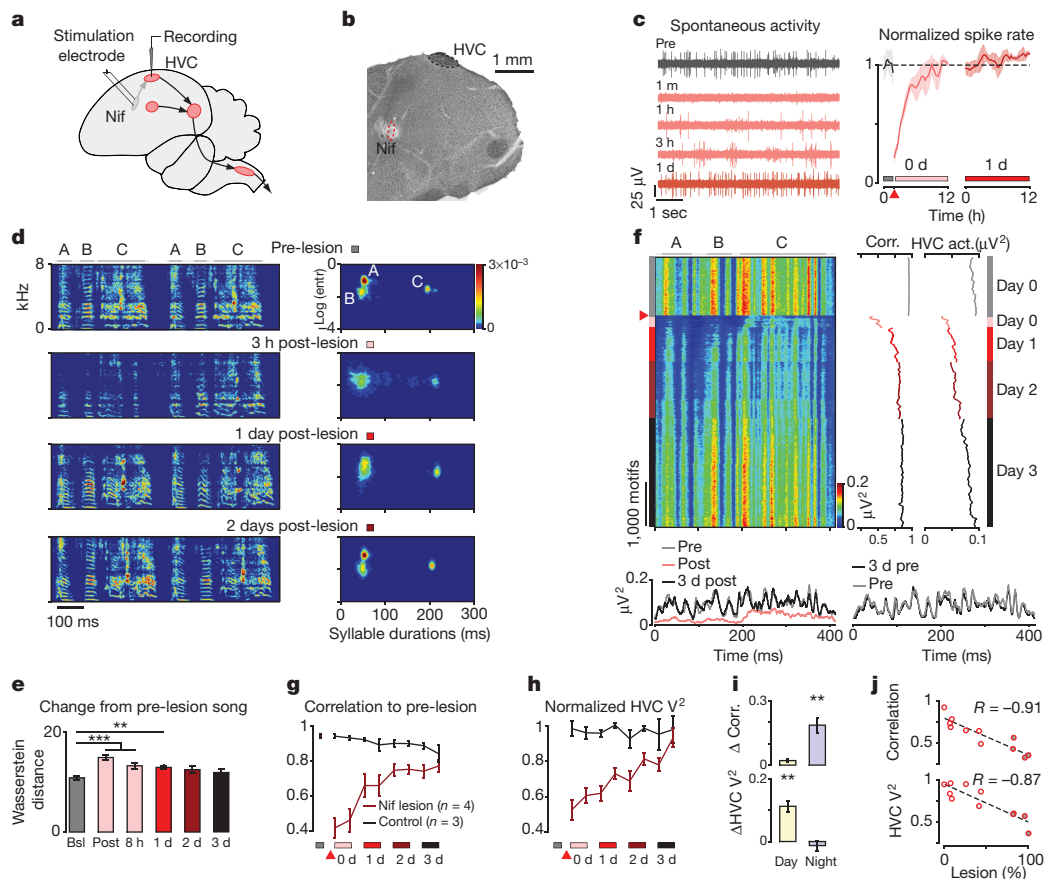
singing. **d**, Schematic showing bilateral Nif inactivation. **e**, Histological sections of Nif. Top, injection of cholera toxin into HVC retrogradely labels Nif (green). Bottom, fluorescent dextran co-injected with muscimol (violet). Red arrows denote estimated boundaries of Nif. **f**, Spectrograms (left) and syllable entropy-duration distributions (right) as in **b** for an example bird subjected to various injection protocols. **g**, Same as **c**, but comparing pre-injection songs to songs after muscimol/PBS injections ( $n = 5$  birds). **h**, Syllable-duration distributions in Nif-inactivated birds compared to HVC-lesioned birds. Data for HVC-lesioned birds from ref. 26. Shown for comparison are the distributions for baseline and Nif inactivation for the bird in **f**. Error bars represent s.e.m. \*\*\* $P < 0.001$ .

song-aligned activity patterns in HVC also recovered their pre-lesion structure (Fig. 3f–h). By the third day, the residual difference was consistent with normal drift in the recordings. Interestingly, the temporal structure of song-aligned HVC activity recovered predominantly during the night, while song-related HVC power recovered during the day (Fig. 3i; Methods).

To assess the extent to which acute post-lesion changes in HVC dynamics were caused by removal of Nif input versus non-specific effects of the current injections, we quantified changes in song-related HVC dynamics as a function of Nif lesion size. We found the extent of Nif damage to be strongly correlated with changes in song-related HVC activity following lesions (Fig. 3j), consistent with the acute degradation of song and associated HVC activity being due to removal of Nif input to HVC.

### Activity homeostasis explains functional recovery

The spontaneous and gradual recovery of HVC activity after Nif lesions (Fig. 3c, h) was suggestive of homeostatic regulation of neural activity<sup>7–9</sup>. To probe whether this could explain the observed song recovery, we modelled HVC as a synaptically connected chain of neurons (a 'synfire chain') that receives time-varying excitatory input from Nif<sup>32,33</sup> (Fig. 4a; Methods). The network generated stable propagation of synchronous spiking activity, much like what is assumed for HVC during singing<sup>34</sup>. Acute removal of Nif input prevented many neurons in the chain from reaching spiking threshold, causing activity propagation to slow and often stop prematurely (Fig. 4c, d). Homeostatic regulation of neural activity in the HVC network was implemented by adaptively adjusting either spiking threshold<sup>35</sup> (Fig. 4b), input resistance<sup>36</sup> or strength of synaptic inputs<sup>37</sup> (Extended Data Fig. 6) of individual HVC neurons (Methods). These mechanisms all had similar effects: increasing the probability of HVC spiking while speeding up chain propagation and decreasing the likelihood of early 'song' terminations (Figs 4b–d and Extended Data Fig. 6).



**Figure 3 | Initial disruption and subsequent recovery in vocal performance and HVC dynamics following Nif lesions.** **a**, Schematic of the experiment: lesioning Nif unilaterally while continuously recording neural activity from ipsilateral HVC. **b**, Histology of an electrolytic Nif lesion. **c**, Left, spontaneous activity in HVC before and at different time points after unilateral Nif lesion in an example bird. Right, recovery of spontaneous HVC activity normalized to pre-lesion rates averaged across Nif lesioned birds ( $n = 4$ ) (lesion indicated by red arrow). Spontaneous activity recovered with a time constant of  $3.4 \pm 1.5$  h. Shaded region denotes s.e.m. **d**, Representative spectrograms (left) and joint entropy-duration distributions (right) for songs of an example bird before and at different times after unilateral Nif lesion. **e**, Summary statistics showing the difference (Wasserstein distance) between the joint entropy-duration probability distributions before, and at different times after, unilateral Nif lesions. Baseline ('Bsl') compares the distributions from two consecutive days of pre-lesion singing. **f**, Recovery in song-related HVC dynamics following Nif lesion for the bird in **d**. Left, song-aligned neural power in HVC for song motifs uttered on the day of Nif lesion (red arrow) and until 3 days after. Middle, correlation between the song-aligned HVC activity pattern and the average pre-lesion activity pattern. Right, average neural power in HVC during singing. Bottom-left panel, average song-aligned

neural activity right before and after the lesion, and 3 days later. Bottom-right panel, same as on the left, but showing normal drift in the neural recording over the 3 days preceding the lesion. **g**, Recovery in song-aligned HVC activity following unilateral Nif lesions measured as the Pearson's correlation to the average pre-lesion activity pattern and averaged across 4 birds (red trace, Methods). Data points correspond to the first and last batch of 25 song motifs on each day. The control trace (black) comes from recordings in 3 of the 4 lesioned birds, but before the Nif lesions, and represents the expected drift in HVC recordings. **h**, Similar to **g**, but showing recovery of the mean neural power in HVC during song, normalized to pre-lesion power. **i**, Recovery in HVC dynamics parsed by day (start to end of day-time singing) and night (end of singing on one day to start of singing the next), over the first 60 h post-lesion. Top, recovery in the correlation to pre-lesion HVC dynamics. Bottom, recovery of mean neural power. **j**, Correlation between song-aligned HVC activity before and immediately after Nif lesions (top), and the mean song-related neural power immediately after lesions normalized to pre-lesion values (bottom), as a function of the fraction of Nif lesioned. Grey-filled circles identify birds with  $>80\%$  Nif lesions that were included in the summary analyses (**c**, **e**–**i**). Error bars represent s.e.m. \*\*\* $P < 0.001$ , \*\* $P < 0.01$ .

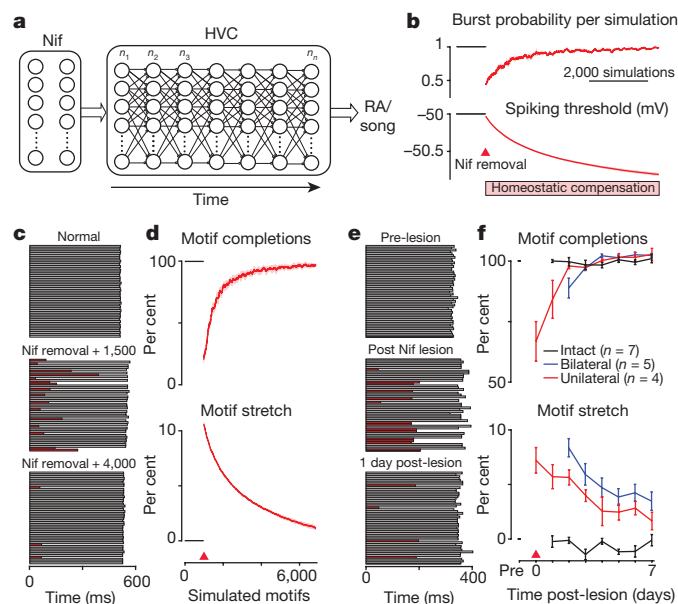
Given that HVC is known to control song timing<sup>25</sup>, our simulations generated predictions for the temporal structure of song following Nif lesion (Fig. 4d). In agreement with these predictions, we found a transient increase in premature song terminations (Fig. 4e, f and Extended Data Fig. 5). Upon closer inspection, we also found that the song slowed down after lesion<sup>18</sup>, only to recover in the ensuing days (Fig. 4f), again consistent with the qualitative predictions of our model. Though we cannot exclude other mechanisms, our network simulations are consistent with functional recovery after Nif lesions being due to homeostatic regulation of neural activity in HVC.

## Discussion

Although efforts to understand the brain must necessarily rely on reductionist approaches, the simplifications and assumptions made

in this pursuit must be scrutinized to prevent misleading conclusions. Given the increased reliance on transient circuit manipulations (for example, optogenetics, pharmacology, pharmacogenetics, cooling and transcranial magnetic stimulation) for localizing brain function, we tested whether behavioural effects induced by sudden activity perturbations reliably reflect the computations carried out in targeted areas. In two different systems, the deficits induced by transient manipulations seemingly overestimated the steady-state function of the examined circuits (Figs 1 and 2). This could be explained by the manipulations acutely affecting the independent functions of downstream circuits. We found that such off-target effects can resolve after the targeted area is permanently lesioned (Fig. 3). Importantly, the post-lesion recovery process did not require any renewed experience with the task, and was consistent with homeostatic regulation of neural activity (Fig. 4).





**Figure 4 | Homeostatic regulation of spiking activity in HVC neurons can account for the functional recovery after Nif lesions.** **a**, HVC is modelled as a chain of synchronously firing neurons, with each neuron receiving a different time-varying input from Nif (Methods). **b**, Top, Nif removal reduces the probability of spiking during simulated 'song' in a model neuron (from the 40th node). Bottom, homeostatic regulation of neural activity is implemented by adaptively adjusting the spiking threshold (Methods). **c**, **d**, Behaviour of the HVC model network after removal of Nif input and subsequent homeostatic regulation of single neuron firing rates. **c**, Fifty simulated 'songs' before and at two different times (1,500 and 4,000 simulations respectively) after Nif removal. Completed 'songs' in grey; truncated ones in red. **d**, Top, fraction of simulations for which activity in the model HVC network propagated to the end. Bottom, average duration of a full chain-propagation as a function of homeostatic recovery. **e**, **f**, Similar to **c** and **d**, respectively, but for birds with unilateral ( $n = 4$ ) and bilateral ( $n = 5$ ) Nif lesions. **f**, Motif completion rates (top) and tempo (bottom) relative to pre-lesion baseline. Data for intact birds come from a subset of the birds that were later lesioned. Error bars represent s.e.m.

Discrepancies between acute and chronic behavioural effects of targeted inactivations/lesions have been recognized in other contexts<sup>10–12,14</sup>. Acute effects are almost invariably more severe, a discrepancy typically explained by the brain adaptively compensating for lost function after lesions<sup>15</sup>. By not allowing time for experience-dependent compensation, transient circuit manipulations are seen as overcoming this 'caveat' of lesions. However, if the goal is to assign computations and memory functions to specific brain areas, our results suggest that transient circuit manipulations may have their own interpretive difficulties that stem from acute effects on the independent functions of non-targeted circuits.

That the function of a circuit can be sensitive to sudden perturbations in chronically non-essential inputs is not surprising. The brain—a finely tuned, complex, and heavily interconnected dynamical system—operates in a fairly limited dynamic regime<sup>38</sup>, making it plausible that local circuit perturbations could interfere with the dynamics and independent functions of remote circuits<sup>39</sup>. For example, sudden removal of permissive inputs could tilt a network's excitatory-inhibitory balance<sup>40</sup>, thus compromising its function<sup>41</sup>. This is seemingly what happens to HVC after Nif is silenced. Loss of excitatory input from Nif causes an acute decrease in the activity of HVC neurons<sup>30</sup>, rendering the network incapable of producing its normal output (Figs 3f, g and 4).

The intricacies of dissecting interconnected biological networks and assigning functions to discrete nodes in those networks have been recognized in other contexts, including genetic and molecular networks<sup>42</sup>. In such studies, the distinction between permissive and instructive

functions is routinely made<sup>43,44</sup>. Our results suggest that a similar distinction should be considered when interrogating the role of neural circuits in behaviour<sup>45</sup>, with a circuit being classified as 'permissive' if its activity is acutely required for the expression of a behaviour without providing essential information for any of the underlying computations or memories. In contrast, a brain area should be considered 'instructive' if it contributes essential information or computation not otherwise available to the system implementing the behaviour.

Although the behavioural effects of sudden activity perturbations may not reliably reflect the steady-state function(s) of a circuit, lesions can, in certain cases at least, contribute additional insight. Permanent silencing of Nif and motor cortex suggested that the capacity of these brain areas to influence the respective skills we study—evident from transient manipulations—is not exercised under normal conditions, consistent with permissive roles. That Nif and motor cortex have access to the essential control circuits likely reflects instructive roles for these brain areas in behavioural processes that we did not test. Nif, for example, provides early auditory priming of HVC essential for imitative song learning<sup>22</sup>, while motor cortical input to subcortical motor circuits is required for the initial acquisition of the skills we train<sup>16</sup> and for modulating other low-level motor behaviours<sup>46</sup>.

Importantly, neural circuit function acutely compromised by sudden changes in permissive input can recover after those inputs are permanently silenced. Both skills we studied recovered after lesions without any task-specific practice, suggesting largely spontaneous recovery processes. Although the mechanisms that underlie such recovery will need to be further examined, our results are consistent with a role for homeostatic regulation of neural activity<sup>7–9</sup> (Fig. 4 and Extended Data Fig. 6). A similar recovery to the one we observed in HVC of songbirds (Fig. 3) has been described for the network underlying the pyloric rhythm in crustaceans<sup>6</sup>, where homeostatic regulation of neuronal dynamics is thought to underlie the recovery of circuit function after removal of permissive or modulatory input<sup>5</sup>.

Interestingly, we found that the structure of song-aligned HVC activity recovered predominantly overnight, while overall HVC power recovered during the day (Fig. 3i). This dissociation is consistent with the synaptic homeostasis hypothesis of sleep<sup>47</sup> that posits synaptic potentiation during wakefulness and synaptic rescaling and memory consolidation during sleep. Our results suggest that sleep not only consolidates activity patterns associated with recent experiences<sup>48</sup>, but may help restore previously established circuit dynamics, and could hence promote functional recovery after brain lesions<sup>49</sup>.

As in our experimental animals, patients with lesions to motor-related brain areas have motor deficits that resolve in the days and weeks following the injury<sup>50</sup>. Aspects of this recovery are thought to be independent of rehabilitation<sup>13</sup>, suggesting spontaneous processes at work. Diaschisis is a broad clinical term referring to the temporary effects of focal brain lesions on remote brain areas<sup>3</sup>, yet the underlying mechanisms remain poorly understood<sup>41</sup>. Our results suggest that focal brain lesions can affect neural dynamics and function in remote brain areas, and that homeostatic regulation of neuronal dynamics may help resolve such acute effects, thus contributing to functional recovery after brain injury.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 17 June; accepted 9 November 2015.**

**Published online 9 December 2015.**

1. Lomber, S. G. The advantages and limitations of permanent or reversible deactivation techniques in the assessment of neural function. *J. Neurosci. Methods* **86**, 109–117 (1999).
2. Zhang, F., Aravanis, A. M., Adamantidis, A., de Lecea, L. & Deisseroth, K. Circuit-breakers: optical technologies for probing neural signals and systems. *Nature Rev. Neurosci.* **8**, 577–581 (2007).
3. Carrera, E. & Tononi, G. Diaschisis: past, present, future. *Brain* **137**, 2408–2422 (2014).

4. Honey, C. J. & Sporns, O. Dynamical consequences of lesions in cortical networks. *Hum. Brain Mapp.* **29**, 802–809 (2008).
5. Golowasch, J., Casey, M., Abbott, L. F. & Marder, E. Network stability from activity-dependent regulation of neuronal conductances. *Neural Comput.* **11**, 1079–1096 (1999).
6. Thoby-Brisson, M. & Simmers, J. Long-term neuromodulatory regulation of a motor pattern-generating network: maintenance of synaptic efficacy and oscillatory properties. *J. Neurophysiol.* **88**, 2942–2953 (2002).
7. Keck, T. *et al.* Synaptic scaling and homeostatic plasticity in the mouse visual cortex *in vivo*. *Neuron* **80**, 327–334 (2013).
8. Marder, E. & Goaillard, J.-M. Variability, compensation and homeostasis in neuron and network function. *Nature Rev. Neurosci.* **7**, 563–574 (2006).
9. Turrigiano, G. G. Homeostatic plasticity in neuronal networks: the more things change, the more they stay the same. *Trends Neurosci.* **22**, 221–227 (1999).
10. Bender, D. B. & Baizer, J. S. Saccadic eye movements following kainic acid lesions of the pulvinar in monkeys. *Exp. Brain Res.* **79**, 467–478 (1990).
11. Wilke, M., Turchi, J., Smith, K., Mishkin, M. & Leopold, D. A. Pulvinar inactivation disrupts selection of movement plans. *J. Neurosci.* **30**, 8650–8659 (2010).
12. Talwar, S. K., Musial, P. G. & Gerstein, G. L. Role of mammalian auditory cortex in the perception of elementary sound properties. *J. Neurophysiol.* **85**, 2350–2358 (2001).
13. Van Peppen, R. P. *et al.* The impact of physical therapy on functional outcomes after stroke: what's the evidence? *Clin. Rehabil.* **18**, 833–862 (2004).
14. Newsome, W. T. & Pare, E. B. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.* **8**, 2201–2211 (1988).
15. Maldonado, M. A., Allred, R. P., Felthaus, E. L. & Jones, T. A. Motor skill training, but not voluntary exercise, improves skilled reaching after unilateral ischemic lesions of the sensorimotor cortex in rats. *Neurorehabil. Neural Repair* **22**, 250–261 (2008).
16. Kawai, R. *et al.* Motor cortex is required for learning but not for executing a motor skill. *Neuron* **86**, 800–812 (2015).
17. Immelmann, K. in *Bird Vocalizations* (ed. Hinde, R.A.) 61–74 (Cambridge Univ. Press, 1969).
18. Cardin, J. A. Sensorimotor nucleus Nlf is necessary for auditory processing but not vocal motor output in the avian song system. *J. Neurophysiol.* **93**, 2157–2166 (2005).
19. Huber, D. *et al.* Multiple dynamic representations in the motor cortex during sensorimotor learning. *Nature* **484**, 473–478 (2012).
20. Peters, A. J., Chen, S. X. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510**, 263–267 (2014).
21. Martin, J. H. Autoradiographic estimation of the extent of reversible inactivation produced by microinjection of lidocaine and muscimol in the rat. *Neurosci. Lett.* **127**, 160–164 (1991).
22. Roberts, T. F., Gobes, S. M. H., Murugan, M., Ölveczky, B. P. & Mooney, R. Motor circuits are required to encode a sensory model for imitative learning. *Nature Neurosci.* **15**, 1454–1459 (2012).
23. Zhang, F., Wang, L.-P., Boyden, E. S. & Deisseroth, K. Channelrhodopsin-2 and optical control of excitable cells. *Nature Methods* **3**, 785–792 (2006).
24. Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nature Methods* **11**, 338–346 (2014).
25. Long, M. A. & Fee, M. S. Using temperature to analyse temporal dynamics in the songbird motor pathway. *Nature* **456**, 189–194 (2008).
26. Aronov, D., Andalman, A. S. & Fee, M. S. A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science* **320**, 630–634 (2008).
27. Fee, M. S. & Scharff, C. The songbird as a model for the generation and learning of complex sequential behaviors. *ILAR J.* **51**, 362–377 (2010).
28. Simpson, H. B. & Vicario, D. S. Brain pathways for learned and unlearned vocalizations differ in zebra finches. *J. Neurosci.* **10**, 1541–1556 (1990).
29. Ali, F. *et al.* The basal ganglia is necessary for learning spectral, but not temporal, features of birdsong. *Neuron* **80**, 494–506 (2013).
30. Hahnloser, R. H. R. & Fee, M. S. Sleep-related spike bursts in HVC are driven by the nucleus interface of the nidopallium. *J. Neurophysiol.* **97**, 423–435 (2007).
31. Schmidt, M. F., Ashmore, R. C. & Vu, E. T. Bilateral control and interhemispheric coordination in the avian song motor system. *Ann. NY Acad. Sci.* **1016**, 171–186 (2004).
32. Long, M. A., Jin, D. Z. & Fee, M. S. Support for a synaptic chain model of neuronal sequence generation. *Nature* **468**, 394–399 (2010).
33. McCasland, J. S. Neuronal control of bird song production. *J. Neurosci.* **7**, 23–39 (1987).
34. Hahnloser, R. H., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419**, 65–70 (2002).
35. Watt, A. J. & Desai, N. S. Homeostatic plasticity and STDP: keeping a neuron's cool in a fluctuating world. *Front. Synaptic Neurosci.* **2**, 5 (2010).
36. van Welie, I., van Hooft, J. A. & Wadman, W. J. Homeostatic scaling of neuronal excitability by synaptic modulation of somatic hyperpolarization-activated  $I_h$  channels. *Proc. Natl Acad. Sci. USA* **101**, 5123–5128 (2004).
37. Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C. & Nelson, S. B. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature* **391**, 892–896 (1998).
38. van Vreeswijk, C. & Sompolinsky, H. Chaotic balanced state in a model of cortical circuits. *Neural Comput.* **10**, 1321–1371 (1998).
39. London, M., Roth, A., Beeren, L., Häusser, M. & Latham, P. E. Sensitivity to perturbations *in vivo* implies high noise and suggests rate coding in cortex. *Nature* **466**, 123–127 (2010).
40. Shu, Y., Hasenstaub, A. & McCormick, D. A. Turning on and off recurrent balanced cortical activity. *Nature* **423**, 288–293 (2003).
41. Feeney, D. M. & Baron, J. C. Diaschisis. *Stroke* **17**, 817–830 (1986).
42. Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev. Genet.* **9**, 855–867 (2008).
43. Miyashita, T., Kubik, S., Lewandowski, G. & Guzowski, J. F. Networks of neurons, networks of genes: an integrated view of memory consolidation. *Neurobiol. Learn. Mem.* **89**, 269–284 (2008).
44. Shobe, J. The role of PKA, CaMKII, and PKC in avoidance conditioning: permissive or instructive? *Neurobiol. Learn. Mem.* **77**, 291–312 (2002).
45. Taha, S. A. & Fields, H. L. Inhibitions of nucleus accumbens neurons encode a gating signal for reward-directed behavior. *J. Neurosci.* **26**, 217–222 (2006).
46. Stoltz, S., Humm, J. L. & Schallert, T. Cortical injury impairs contralateral forelimb immobility during swimming: a simple test for loss of inhibitory motor control. *Behav. Brain Res.* **106**, 127–132 (1999).
47. Tononi, G. & Cirelli, C. Sleep function and synaptic homeostasis. *Sleep Med. Rev.* **10**, 49–62 (2006).
48. Walker, M. P. & Stickgold, R. Sleep-dependent learning and memory consolidation. *Neuron* **44**, 121–133 (2004).
49. Siccoli, M. M., Röhl-Baumeler, N., Achermann, P. & Bassetti, C. L. Correlation between sleep and cognitive functions after hemispheric ischaemic stroke. *Eur. J. Neurol.* **15**, 565–572 (2008).
50. Levin, H. S. & Grafman, J. *Cerebral Reorganization of Function after Brain Damage* (Oxford Univ. Press, 2000).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank E. Soucy and J. Greenwood for technical assistance. We are grateful to M. Meister, J. Sanes, N. Uchida, K. Blum, A. Dhawale, M. Josch, A. Kampff, and E. Feinberg for their feedback on our manuscript. This work was supported by a McKnight Scholars Award to B.P.Ö., HSFP and EMBO fellowships to S.B.E.W., an NRSA fellowship to R.K., and a Rubicon fellowship from the Netherlands Organization for Scientific Research to S.M.H.G.

**Author Contributions** B.P.Ö. and T.M.O. designed the study with input from all authors. T.M.O. collected and analysed the data from songbirds with help from A.K. S.M.H.G. did initial pilot experiments in songbirds that inspired the study. C.P. implemented the HVC network model. S.B.E.W. performed the optogenetics experiments in rats and analysed the data. J.Y.R. and R.K. performed the pharmacological inactivation experiments in rats and analysed the data. B.P.Ö. supervised and coordinated the project. B.P.Ö., T.M.O., and S.B.E.W. wrote the paper with input from the other authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.P.Ö. ([olvecky@fas.harvard.edu](mailto:olvecky@fas.harvard.edu)).

## METHODS

**Animals.** The care and experimental manipulation of all animals were reviewed and approved by the Harvard Institutional Animal Care and Use Committee. Experimental subjects were female Long Evans rats 3–8 months old at start of training ( $n = 10$ , Charles River) and adult male zebra finches between 92–205 days post-hatch ( $n = 27$ ). Because the behavioural effects of our circuit manipulations could not be pre-specified before the experiments, we chose sample sizes that would allow for identification of outliers and for validation of experimental reproducibility. No animals were excluded from experiments post-hoc. The investigators were not blinded to allocation during experiments and outcome assessment, unless otherwise stated.

**Behavioural training in rats.** Ten rats were trained in the lever-pressing task as previously described<sup>16</sup>. Water-restricted animals were rewarded with water for pressing a lever twice with a prescribed interval between the presses (700 ms for 9 of the rats, 600 ms for one). All animals were trained using our fully automated home-cage training system<sup>51</sup>. Kinematic tracking of forepaw movements (Fig. 1d, g) was done as in ref. 16.

**Motor cortex inactivations in rats.** In rats ( $n = 5$ ) that had reached asymptotic performance in our task<sup>16</sup>, a craniotomy was made to access the caudal forelimb area of primary motor cortex (CFA) in the hemisphere contralateral to the paw most involved in the lever-press sequence. The centre of the CFA was estimated from stereotactic coordinates (+1.0 mm anterior, +3.00 mm lateral, with respect to bregma<sup>52</sup>). Kwik-Kast sealant (WPI) was applied to cover the exposed dura. In addition, a protective acrylic cap covering the craniotomy was attached with screws to three nuts secured to the skull with Metabond (Parkell). After recovering from surgery (10 days), animals were trained for at least one additional week to ensure that they were at their asymptotic performance levels.

On injection days, rats were lightly anaesthetized with 0.5–1.5% isoflurane and placed in a stereotax. Motor cortex was accessed by removing the custom-made protective cap and the Kwik-Kast plug covering the craniotomy. Muscimol (or PBS for control) was injected at the estimated centre of the CFA<sup>52</sup>, 1.5 mm deep, in 9.2 nl increments every 10 s using a Nanoject (WPI). The craniotomy was resealed with Kwik-Cast and the protective cap reattached. The whole procedure took 15–30 min, and rats resumed normal behaviour a few minutes later. Training sessions started 1.5 h after the injections. ‘Baseline’ performance included sessions after the craniotomies but before injection. Experimental days alternated between saline and muscimol injections. To prevent any behavioural compensation in response to muscimol-induced performance deficits, injected animals were tested for only 10 min.

The dosing of muscimol was based on two criteria. (1) To allow comparisons with our lesion study<sup>16</sup>, the direct effect of muscimol injections should be restricted to a volume of motor cortex equal to or smaller than what we lesioned in ref. 16 (Fig. 1c). (2) To quantify effects on kinematics and performance, drug dosing should not abolish task engagement. We injected increasing concentrations and volumes of muscimol in two rats, and found that relatively larger doses (200–400 nl of 25 mM muscimol) degraded performance to the point where animals quit the task (Supplementary Video 1).

We converged on a dose of 100 nl of 25 mM muscimol because it generally did not prevent engagement with the task and because we estimated that it affects a volume significantly smaller than what we lesioned in ref. 16. Our estimate of muscimol spread is bounded by previous studies that injected larger doses of muscimol<sup>21,53</sup> (1  $\mu$ l of 2 and 9 mM, respectively) and showed affected volumes of ~4–14 mm<sup>3</sup>. In comparison, our motor cortex lesions were larger than 23 mm<sup>3</sup> (ref. 16). We also injected 100 nl of 25 mM fluorescein in one animal using the same protocol as for the experimental animals, euthanizing it 1.5 h after the injection. Its brain was later sectioned, and the approximate spread of the dye visualized using fluorescence microscopy (Fig. 1c). One of the experimental animals had very severe performance deficits at our chosen dosing, preventing us from characterizing the behavioural effect (that is, very few successful trials). In this animal we reduced the concentration to 1 mM, at which the task engagement was robust but the performance still affected (Fig. 1d).

**Optogenetic stimulation.** *Viral injections.* Adeno-associated virus (AAV2/8-hSyn-FLEX-ChrimsonR-tdTomato, UNC vector core<sup>24</sup>; titre:  $5 \times 10^{12}$  vector genomes (vg) per ml) was injected into the forelimb motor cortex of isoflurane anaesthetized rats ( $n = 5$ ) through multiple small craniotomies (A/P, M/L: +1,  $\pm 2$ ; +1,  $\pm 4$ ; +1.5,  $\pm 2.75$ ; +2.25,  $\pm 2.5$ ; +3,  $\pm 2$ , coordinates relative to bregma). Injections were done in 9.2 nl increments while slowly moving the injection-pipette (Nanoject) from a depth of 1.5 mm to 0.7 mm for a total volume of 0.4  $\mu$ l per site and 2  $\mu$ l per hemisphere (Extended Data Fig. 1a). Animals were allowed to recover for 5 days before starting behavioural training.

**LED implant and stimulation.** Once animals reached asymptotic performance on our task<sup>16</sup>, they underwent a second surgery to implant a custom-built device for optogenetic stimulation (Extended Data Fig. 1c). The device consisted of a red

light-emitting diode (LED) ( $\lambda = 615$  nm, 110 mW output power, XLAMP XPC LED RD-ORANGE, Cree) on a printed circuit board, powered by two coin-cell batteries (CR2032). An infrared (IR)-sensitive photodiode was used to wirelessly control the LED. After device implantation and recovery, animals resumed behavioural training. An IR light-source placed on top of the training cage was activated to trigger the LED for the duration of the optogenetic stimulation (1 s or 50 ms, continuous light). Stimulation trials were at least 10 s apart to allow the batteries of the LED to recover. Between stimulation trials, rats performed a varying number of non-stimulated trials (range: 0–5), resulting in ~30% of the trials being ‘stimulated’. Optogenetic stimulation was repeated for several sessions (5–12). Batteries were changed daily.

**Functional verification.** To characterize the effects of optogenetic stimulation on motor cortex activity, acute electrophysiological recordings were done after termination of the behavioural experiments in two of the rats. The animals were anaesthetized and placed in a stereotactic frame. The implanted LED device was carefully removed to expose the previous craniotomy. Using a custom-built recording setup and silicon probes (Buzsaki-64, Neuronexus), we recorded single-unit activity in motor cortex below the craniotomy. The removed LED device was placed next to the silicon probe above the craniotomy. Once stable units were detected, we triggered the LED and illuminated motor cortex for 1 s (30 trials). Recordings were performed at multiple depths (0.1 mm to 2.5 mm). Units were classified as light responsive if at least two consecutive bins of 5 ms during the first 200 ms of illumination had a significant  $z$ -score (compared to 1 s of baseline before light onset). These included units with long onset latencies (>10 ms), consistent with indirect activation (Extended Data Fig. 1b). The relatively high number of light responsive units ( $69 \pm 3\%$ ), compared to the number of cells counted as infected by immunohistochemistry  $31 \pm 2\%$ ; see below), is likely due to such indirect effects. Moreover, many of the recorded light-responsive cells were only identified during stimulation, further biasing our results to responsive cells.

**Histological verification.** At the end of the experiments, animals were transcardially perfused with PBS and subsequently fixed with 4% paraformaldehyde (PFA) in PBS. Brains were removed and post-fixed for at least 24 h. Brains were sliced coronally (thickness: 80  $\mu$ m) and immunohistochemistry performed to determine the AAV injection site and extent of the transfection (Extended Data Fig. 1a). Slices were blocked (1% BSA, 0.3% Triton in PBS) at room temperature and incubated with anti-RFP (chicken, 1:1,000, Millipore, AB3528) and anti-NeuN (mouse, Millipore, MAB377) primary antibodies in blocking buffer for 48 h at 4 °C. After washing, slices were incubated with anti-chicken-Alexa 568 (goat, 1:1,000, Life Technologies, A-11041) and anti-mouse-Alexa 647 (goat, 1:1,000, Life Technologies, A-31625) over night at 4 °C. Slices were mounted and imaged using a Zeiss Axio Scan Z1 Slide Scanner for overview images and an Olympus FluoView FV1000 confocal microscope for high-resolution images. We verified the targeting and spread of all injections based on the fluorescent signal and determined the extent of the AAV injections in a subset of animals ( $8.2 \text{ mm}^3 \pm 1.3 \text{ mm}^3$ ;  $n = 2$  rats). In addition we chose 4 regions of interest (size  $635 \times 635 \mu\text{m}$ ) and counted the number of infected cells relative to the number of neurons (NeuN<sup>+</sup> cell) to determine the fraction of infected cells ( $31 \pm 2\%$ ;  $n = 2$  rats). Histology was done blind to the outcome of the experiment.

**Data analysis for rat experiments.** To assess the behavioural effects of the different injections in the acute inactivation experiments, we measured performance relative to ‘baseline’ training sessions after the craniotomies but before any injections. To standardize analysis across experimental conditions (muscimol, PBS, or baseline), we only included data from the first 10 min of each session, matching the duration of the muscimol sessions. For the optogenetic stimulation experiments, ‘baseline’ was defined as the non-stimulated trials in the same sessions. The number of sessions for each condition ranged from 1–3 (injections) and 5–12 (optogenetic stimulations). Data from training sessions of a given condition were pooled for each animal. To quantify behavioural performance, the fraction of trials with an inter-press interval (IPI) within 20% of the target IPI was calculated. Data on motor cortex lesioned animals presented for comparisons in Fig. 1e comes from previously published experiments<sup>16</sup>, and includes sessions from the second week of post-lesion training.

**Zebra finch experiments.** All birds were obtained from the Harvard University zebra finch breeding facility and housed on a 13:11 h light/dark cycle in acoustic isolation with food and water provided ad libitum.

**Pharmacological lesions.** Birds ( $n = 5$ ) were anaesthetized with isoflurane. Nif was localized antidromically by electrical stimulation in HVC<sup>29</sup>. Bilateral Nif lesions were made by injecting the excitotoxin *N*-methyl-DL-aspartic acid (NMA, 4%) into each hemisphere using a Nanoject (WPI). In initial experiments, a single 27 nl bolus of NMA was injected into the centre of Nif. Though this volume produced complete bilateral Nif lesions in one animal, we found that complete lesions were more reliably produced by injecting two boluses of 18 nl (for a total of 36 nl) 200  $\mu$ m apart along the anterior-posterior axis. We report on the five animals that had 100%



bilateral Nif lesions, determined by post-hoc histological inspection (see below, Extended Data Fig. 3b). One of these received 27 nl and 4 received 36 nl injections. **Reversible inactivations.** Birds ( $n = 5$ ) were anaesthetized and Nif identified as described above. Craniotomies over Nif were covered with artificial dura (Body Double Fast; Smooth-On, Inc.) and head screws were attached to the skull with dental cement as previously described<sup>54</sup>. Following post-surgical recovery, awake birds were placed in a foam restraint and head-fixed to a stereotax for ~10 min each morning for 10–14 days to desensitize them to handling and restraint. Following the desensitization training, all birds reliably sang within 30 min of the restraint. In the morning of experimental days, muscimol (27 nl, 50 mM) or PBS (27 nl) was injected bilaterally as described in the text. In two birds that routinely sang within 10 min of drug administration, we also injected a smaller dose (9 nl) of muscimol into Nif to additionally verify that song degradation was due to direct inactivation of Nif (Extended Data Fig. 4b).

We note that a previous study aimed at reversibly inactivating Nif in adult songbirds failed to show any obvious effect on song structure<sup>55</sup>, but conflicting results from experiments in juvenile birds and methodological uncertainties regarding drug injection volumes in adult birds make its conclusions tentative. This previous report notwithstanding, all our muscimol injections into Nif produced similarly severe song degradation (Fig. 2f and Extended Data Fig. 4b).

**Implantation of recording and stimulation device.** Zebra finches ( $n = 11$ ) were anaesthetized with isoflurane and placed in a stereotax. HVC was identified by antidromic stimulation from Area X as previously described<sup>29</sup>. Nif was similarly identified by stimulating in HVC. For birds targeted for electrolytic Nif lesions, we placed either a monopolar stimulating electrode at the dorsal-posterior edge of Nif ( $n = 8$  birds) or a bipolar stimulating electrode straddling Nif in the medial-lateral plane ( $n = 3$  birds). A custom recording array (3 channels; 100 k $\Omega$ ) was implanted in the hemisphere ipsilateral to the Nif-stimulating electrode and within the identified boundaries of HVC as previously described<sup>29</sup>. All birds exhibited normal song output within 7 days of surgery. Following completion of the experiment, animals were euthanized, their brains collected, and the placement of recording electrodes and extent of lesions confirmed histologically.

**Neural and behavioural recordings.** Sound and neural activity were recorded using a custom LabVIEW application (National Instruments) as previously described<sup>29</sup>. Multi-unit neural activity was recorded from up to three sites in HVC (~250  $\mu$ m spacing) for three to four weeks per bird. Because stability of the neural recordings is crucial for estimating recovery in HVC dynamics, analysis was done on data collected at the most stable recording site in each bird (determined pre-lesion), though we note that the trends were similar across all channels.

**Electrolytic lesions.** Electrolytic lesions of Nif were made in the right hemisphere by passing 50  $\mu$ A of monophasic current through the stimulating electrode for 30–40 s. Current injections started while birds were singing, and in all cases immediately terminated song output. Lesion extent was estimated post-hoc as described below.

**Histological verification of lesions and inactivations.** At the end of the experiments, birds were anaesthetized with natriumpentobarbital (Nembutal, IM) and transcardially perfused with PBS, followed by fixation with 4% PFA in PBS. Brains were removed and post-fixed in 4% PFA overnight. Parasagittal sections (75  $\mu$ m) were cut on a Vibratome (Leica), mounted, and stained with cresyl violet to reconstruct the location of implanted electrodes and lesions (ImageJ). Identification of the injection sites for the muscimol inactivations and circuit tracings were done in alternate brain slices by fluorescence microscopy (Fig. 2e and Extended Data Fig. 4a). Histology was done blind to the identity of the animals.

Nif was identified based on regions of stronger staining and higher cell density than surrounding areas and were additionally guided by proximate anatomical landmarks (for example, HVC, the lamina mesopallialis and the lamina pallio-subpallialis).

**Lesions.** Location and size of the lesions were determined by estimating the extent of necrotic tissue (that is, loss of neurons and gliosis) in photomicrographs of cresyl violet stained sections as previously described<sup>22</sup>. Lesion size was expressed as a percentage of estimated Nif size, measured in intact controls ( $0.035 \pm 0.001$  mm<sup>3</sup>,  $n = 4$  birds). In pharmacologically lesioned birds, 100% of Nif was lesioned (Extended Data Fig. 3b). In electrolytically lesioned birds, 0–100% of Nif was lesioned (Fig. 3j).

**Inactivations.** Fluorescent dye-conjugated dextrans (0.5 mg ml<sup>-1</sup> Alexa 594; Invitrogen) were co-injected with the final injection of muscimol for post-hoc verification of the injection site (Fig. 2e and Extended Data Fig. 4a). Fluorescence images of the sections were superimposed on those of their adjacent cresyl violet sections (Adobe Photoshop) to determine locations of fluorescence in relation to Nif. All injection sites were found to be within the target nucleus (Extended Data Fig. 4a).

**Neural circuit tracing.** To visualize Nif (Fig. 2e and Extended Data Fig. 3a), fluorescent dye-conjugated cholera toxin subunit B (1 mg ml<sup>-1</sup>, Alexa 488; Invitrogen) was injected into HVC in 2 birds (83 nl per hemisphere). Twenty-one days after surgery, the animals were euthanized, perfused, and their brains fixed, sectioned,

and mounted. Photomicrographs of fluorescent sections were overlaid on those of adjacent cresyl violet sections (Adobe Photoshop) to determine the location of fluorescence in relation to anatomical landmarks and density of cell bodies.

**Data analysis of song. Syllable segmentation and annotation.** Raw audio recordings were segmented into syllables as previously described<sup>29</sup>. Spectrograms were calculated for all prospective syllables, and a neural network (5,000 input layer, 100 hidden layer, 3–10 output layer neurons) was trained to identify syllable types using a test data set created manually by visual inspection of song spectrograms. Accuracy of the automated annotation was verified by visual inspection of a subset of syllable spectrograms.

**Syllable feature quantification.** All non-call vocalizations were characterized by their duration and mean Wiener entropy—both robust acoustic features that are tightly controlled in adult zebra finch song<sup>56</sup>. Syllable durations were estimated from threshold crossings of the acoustic power as previously described<sup>29</sup>. Wiener entropy, a measure of acoustic randomness, was calculated using Sound Analysis for MATLAB<sup>57</sup> for 10 ms time windows, advancing in steps of 1 ms, such that entropy was computed for every millisecond. The entropy measurements were averaged across the syllable and log-transformed. On this scale, the Wiener entropies of white noise and of a pure tone are zero and minus infinity, respectively.

**Duration probability distributions.** Histograms (1.25 ms bins) of syllable durations produced within 1 h of muscimol/PBS injections were generated for each experiment, normalized by total sample counts, averaged across 2–4 experiments within a bird, and then averaged across birds. Data from HVC-lesioned birds, provided by the authors of ref. 26, was recorded on the first day of singing after lesion (2–7 days after surgery) and analysed similarly. Mean duration distributions for all conditions were smoothed with a sliding boxcar window (7-bin width, 1-bin advance).

**Entropy-duration joint probability distributions.** Two-dimensional histograms, showing the joint distributions of syllable duration and Wiener entropy, were created with bins of width 1.25 ms (duration axis; range: 0–300 ms) and 0.025 (log Wiener entropy axis; range: -4–0). The histogram was normalized by total sample counts to construct an empirical probability distribution. Because these empirical distributions were sparsely sampled, we estimated the true probability distribution by smoothing the empirical distribution with a point-spread function (2D Gaussian; width: 7 bins; sigma: 3 bins). Distributions were calculated for vocalizations produced during the following time windows. Bilateral lesion experiments: the first 2 h of singing each day; inactivation experiments: 2 h before (pre), 1 h after (post), and 6–8 h after (washout) injection; unilateral lesion: the first 2 h of singing each day, the first hour of post-lesion singing, and the last 4 h of singing on the day of lesion.

**Distribution similarity measurement.** To quantify changes in song elements, we calculated the first Wasserstein distance, a common metric of the difference in probability distributions, between syllable entropy-duration distributions for songs produced at different time points or under various experimental conditions (see text). We used an implementation in MATLAB and C available at (<http://www.ariel.ac.il/sites/ofirpele/FastEMD/>). Distances between bins were Euclidean. Calculations were based on 50,000 samples drawn from the entropy-duration probability distributions and reported in figures as the mean distance per sample. **Motif completion rate.** For each bird, a 3–5 syllable dominant song motif was identified by visual inspection of spectrograms. Motif completion rates (MCR) were calculated as:

$$\text{MCR} = \frac{\text{Number of utterances of complete motifs}}{\text{Number of utterances of the first syllable in the motif}}$$

For all birds, motif completion rates were calculated for the first two hours of singing per day; for unilaterally lesioned birds, rates were also calculated for the first hour of singing following lesion. 'Intact' motif completion rates (Fig. 4f) were based on a subset of the lesioned birds (four from the 'bilateral'; three from 'unilateral' group) but collected 1–2 weeks before the Nif lesions. Data from each bird was normalized to pre-lesion motif completion rates for comparison across animals. See Extended Data Fig. 5 for examples of truncated motifs.

**Motif duration stretch.** The durations of the dominant song motifs were calculated as previously described<sup>29</sup> for interval durations. For all birds, the mean motif duration was calculated for 100 consecutive renditions, taken at the same time each day (~1 h after lights on in the morning). For unilaterally lesioned birds, the mean duration was also calculated for the first 100 identifiable motifs produced immediately after lesion. As noted above, 'intact' data were collected from birds that were later lesioned. Motif durations were normalized to pre-lesion values for comparison across animals. See Extended Data Fig. 5 for examples of aligned and excluded vocalizations.

**Data analysis of neural recordings. Spontaneous activity.** To record spontaneous HVC activity, minute long recordings were made every 15 min. These recordings

were bandpass filtered (1–5 kHz; 2-pole Butterworth; zero-phase) and segments within 500 ms of vocalization-related activity were marked for exclusion from subsequent analysis. Individual spikes were detected by an amplitude threshold set to 3–8 standard deviations of the estimated noise in the recordings. For each bird, the spontaneous firing rates were normalized to the mean firing rate in the two hours before lesion. Shown in Fig. 3c is the across-bird mean and standard error, smoothed with a sliding boxcar window (5 bin width, 1 bin advance).

**Alignment of the neural recordings to song.** A dynamic time warping (DTW) algorithm was used to align individual song motifs to a common template as previously described<sup>29</sup>. The warping path derived from this alignment was then applied to the corresponding HVC recordings with a premotor lead of 35 ms<sup>29</sup>. The aligned neural traces were squared (to calculate signal power) and smoothed (5 ms boxcar window, 1 ms advance).

**HVC activity correlation.** The recovery of temporal dynamics in HVC was calculated as the Pearson's correlation between the song-aligned neural power immediately before lesion and the same at different times after lesion. The running correlation in Fig. 3f shows Pearson's correlation between the mean song-aligned activity pattern of pre-lesion songs on the day of lesion and the mean activity patterns in a sliding window of 25 song motifs. The pre-lesion data point in Fig. 3g represents the correlation between the mean power envelopes for two consecutive blocks of 25 motifs recorded immediately before lesion. Normal drift in the song-related HVC signal ('control') was calculated similarly.

**HVC mean power.** The mean HVC power was calculated per motif and averaged over the 25-motif windows as described above for the correlation. For analyses pooled across birds, mean HVC power was normalized to the pre-lesion value.

**Day versus night recovery.** Recovery of HVC activity in the first 60 h following lesion, during which most of the post-lesion recovery occurred, was parsed into 3 day-time and 2 night-time intervals. Daytime recovery was calculated as the change in correlation to pre-lesion activity (or normalized mean power) between the first 25 motifs in the morning (or immediately following lesion) and the last 25 motifs that evening; night-time recovery is the change between the last 25 motifs of the day and the first 25 of the subsequent morning.

**Modelling. Network architecture.** On the basis of previous experimental findings<sup>32,34</sup>, we modelled the HVC network as a synfire-chain of bursting neurons. The model consisted of 1,200 integrate-and-burst neurons organized into 80 nodes. Each of the 15 neurons in a node projected to all neurons in the next node, forming a chain topology.

The subthreshold membrane potential of the  $i^{\text{th}}$  neuron,  $V_i$ , obeys:

$$C \frac{dV_i}{dt} = -g_L (V_i - V_L) + I_{\text{syn},i} + I_{\text{Nif},i} + \sqrt{\tau_\eta} \sigma \eta_i(t)$$

where  $C = 1 \mu\text{F}/\text{cm}^2$  is the membrane capacitance,  $g_L = 0.1 \text{ mS}/\text{cm}^2$  is the leak conductance,  $V_L = -60 \text{ mV}$  is the leak potential,  $I_{\text{syn},i}$  is the synaptic input,  $I_{\text{Nif},i}$  represents external input to the HVC neurons from Nif,  $\eta_i(t)$  is a zero-mean Gaussian white noise with covariance  $\langle \eta_i(t) \eta_i(t') \rangle = \delta(t - t')$ ,  $\tau_\eta = 10 \text{ ms}$  and  $\sigma = 200 \text{ nA}/\text{cm}^2$ . The synaptic input is given by  $I_{\text{syn},i}(t) = W \sum_j M_{ij} \sum_{k, t_j^k < t} \varepsilon(t - t_j^k)$ , where  $t_j^k$  denotes the  $k^{\text{th}}$  spike of  $j^{\text{th}}$  neuron,  $\varepsilon(t) = \Theta(t) e^{-t/\tau_s}$  with  $\Theta(t)$  being the step function and  $\tau_s = 5 \text{ ms}$ ,  $W = 87 \text{ nA}/\text{cm}^2$  and  $M_{ij}$  is 1 for synapses from a neuron  $j$  to the neurons  $i$  in the next node and 0 otherwise. The Nif input is a different waveform for each HVC neuron and does not change across simulations. The waveforms were randomly generated by simulating an Ornstein-Uhlenbeck process with an autocorrelation time scale of 50 ms, starting from a random initial point. Noise and drift were chosen such that the resulting waveforms had a mean of  $97 \text{ nA}/\text{cm}^2$  and standard deviation of  $53 \text{ nA}/\text{cm}^2$ . When the membrane potential of the integrate-and-burst neuron reaches threshold,  $V_{\text{th}} = -50 \text{ mV}$ , the neuron emits 4 spikes with 2 ms intervals, modelling the bursts generated by calcium spikes in RA-projecting HVC neurons<sup>32,58</sup>, and the membrane potential is reset to  $V_R = -55 \text{ mV}$  after a refractory period of 4 ms. Chain propagation was started by a 5 ms pulse input with magnitude  $6.7 \mu\text{A}/\text{cm}^2$  to the neurons in the first node. The parameters of the model were chosen to approximate the results of our experiments. Some of these parameters were subject to change as explained below.

**Homeostatic regulation of neural activity.** We implemented three different homeostatic plasticity rules, each of which can adaptively modify the excitability of HVC neurons.

**Rule 1:** if during a simulated chain propagation a neuron did not spike, its spiking threshold decreased by  $1 \mu\text{V}$ . If the neuron produced more than 8 spikes or 2 bursts, the threshold increased by  $1 \mu\text{V}$ . This rule is used in Fig. 4 and Extended Data Fig. 6a. Such homeostatic changes in spiking thresholds have been observed in experiments<sup>35</sup>.

**Rule 2:** if during a simulated chain propagation a neuron did not spike, the leak conductance of the neuron decreased by  $0.1 \mu\text{S}/\text{cm}^2$ . If the neuron produced

more than 8 spikes or 2 bursts, the leak conductance increased by  $0.1 \mu\text{S}/\text{cm}^2$ . This rule amounts to changing the neuron's input resistance, defined as the change in membrane potential in response to injected current, divided by the current. This rule is used in Extended Data Fig. 6b. Homeostatic changes to input resistance have also been observed in experiments<sup>36</sup>.

**Rule 3:** if during a simulated chain propagation a neuron did not spike, all synaptic weights to that neuron increased by  $6.7 \text{ pA}/\text{cm}^2$ . If the neuron produced more than 8 spikes or 2 bursts, the synaptic weights decreased by  $6.7 \text{ pA}/\text{cm}^2$ . This rule is used in Extended Data Fig. 6c. Activity-dependent homeostatic changes to a neuron's synaptic inputs have been observed in experiments, for example, in cortical neurons<sup>37</sup>.

In Fig. 4 and Extended Data Fig. 6, a 'motif' was considered complete if at least one neuron in each of the 80 nodes produced a spike. Motif duration was calculated as the time from the propagation initiation until the average spike time of the neurons in the last node. We ran simulations with modified parameters to verify that our results presented in Fig. 4 were qualitatively robust.

**Statistical analysis.** All statistics on data pooled across animals is reported in the main text as mean  $\pm$  s.d. and depicted in figure error bars as mean  $\pm$  s.e.m. Where appropriate, distributions passed tests for normality (Kolmogorov-Smirnov), equal variance (Levene), and/or sphericity (Mauchly), unless otherwise noted. Multiple comparison corrected tests were used where justified. Statistical tests for specific experiments were performed as described below.

**Fig. 1e.** Comparison of fraction of trials with IPIs within 20% of the target for different experimental treatments ( $n = 5$  rats). Mauchly's test indicated a violation of sphericity ( $W = 0.134$ ,  $P = 0.049$ ), and a Huynh-Feldt degrees of freedom correction was applied. Subsequent repeated-measures ANOVA revealed significant differences between the treatments ( $F_{(1,17,4,59)} = 35.7$ ,  $P = 0.002$ ). Post-hoc comparisons using Dunnett's test showed significant differences between PBS (control) and muscimol injections ( $P = 0.0002$ ), but not between PBS and baseline ( $P = 0.99$ ).

**Fig. 1h.** Effect of optogenetic stimulation of motor cortex on task performance. A two-tailed, paired  $t$ -test revealed significant differences in performance in the light off and light on conditions ( $n = 5$  rats;  $P = 3 \times 10^{-5}$ ).

**Fig. 2c.** Comparison of Wasserstein distances between joint entropy-duration distributions before and after bilateral Nif lesions ( $n = 5$  birds). Repeated-measures ANOVA showed no significant difference on any day ( $F_{(3,12)} = 2.21$ ,  $P = 0.14$ ).

**Fig. 2g.** Same as Fig. 2c, but comparing pre-injection songs to songs after muscimol/PBS injections ( $n = 5$  birds). Mauchly's test indicated a violation of sphericity ( $W = 1.9 \times 10^{-4}$ ,  $P = 0.014$ ), and a Huynh-Feldt degree of freedom correction was applied. Subsequent repeated-measures ANOVA revealed significant differences between the treatments ( $F_{(1,36,5,43)} = 19.7$ ,  $P = 0.004$ ). Post-hoc comparisons using Dunnett's test showed significant differences between PBS (control) and muscimol injections ( $P = 1 \times 10^{-5}$ ); no other condition significantly differed from PBS ( $P > 0.92$ ).

**Fig. 2h.** Comparison of the syllable duration distributions following HVC lesions ( $n = 5$  birds) and Nif inactivations ( $n = 5$  birds). A Kolmogorov-Smirnov test on the mean distribution across animals showed no significant differences ( $P = 0.24$ ).

**Fig. 3e.** Comparison of Wasserstein distances between joint entropy-duration distributions before and after unilateral lesions to Nif ( $n = 4$  birds). Repeated-measures ANOVA revealed that lesions produced significant differences in song structure ( $F_{(5,15)} = 17.7$ ,  $P = 8 \times 10^{-6}$ ). Post-hoc comparisons using Dunnett's test showed significant differences from baseline until the second day after lesion (post and 8 h:  $P < 0.001$ ; 1 day:  $P = 0.002$ ;  $P > 0.05$  thereafter).

**Fig. 3g.** Comparisons of HVC dynamics in intact controls and following Nif lesions. A two-tailed, paired  $t$ -test revealed significant differences in correlation immediately before and after lesion ( $n = 4$  birds;  $P = 0.003$ ). In addition, two-tailed unpaired  $t$ -tests showed significant differences between lesion and control conditions at matched time points until the third day post-lesion ( $P < 0.03$  before,  $P = 0.1$  at the start of the third day).

**Fig. 3h.** Comparisons of normalized HVC activity in intact controls and following Nif lesion. A two-tailed, paired  $t$ -test revealed significant differences in activity immediately before and after lesion ( $n = 4$  birds;  $P = 0.002$ ). In addition, two-tailed unpaired  $t$ -tests showed significant differences between lesion and control conditions at matched time points until the third day post-lesion ( $P < 0.03$  before,  $P = 0.29$  at the end of the third day).

**Fig. 3i.** Top, comparison of recovery of correlation to pre-lesion HVC dynamics during day and night ( $n = 4$  birds). Two-tailed one-sample  $t$ -tests revealed significant recovery overnight but not during the day (test against mean zero;  $P = 0.01$  and  $P = 0.053$ , respectively). Bottom, comparison of recovery of HVC activity to pre-lesion levels during day and night ( $n = 4$  birds). Two-tailed one-sample  $t$ -tests revealed significant recovery during the day but not overnight (test against mean zero;  $P = 0.007$  and  $P = 0.48$ , respectively).

Fig. 3j. Top, correlation to pre-lesion HVC dynamics immediately following Nif lesions as a function of the fraction of Nif lesioned ( $n = 11$  birds). A two-tailed  $t$ -test revealed the Pearson's linear correlation coefficient,  $R = -0.91$ , to be significantly different from zero ( $P = 1 \times 10^{-4}$ ). Bottom, normalized HVC activity immediately following Nif lesions as a function of the fraction of Nif lesioned ( $n = 11$  birds). A two-tailed  $t$ -test revealed the Pearson's linear correlation coefficient,  $R = -0.87$ , to be significantly different from zero ( $P = 5 \times 10^{-4}$ ).

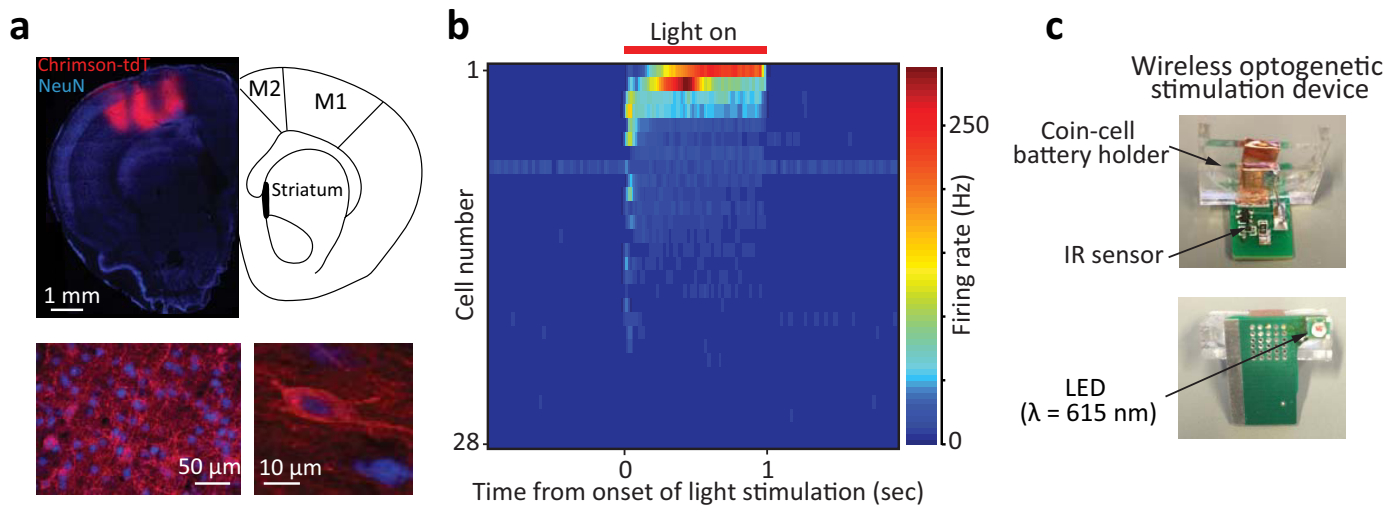
Fig. 4f. Top, comparison of post-lesion motif completion rates to pre-lesion baseline for unilateral ( $n = 4$  birds) and bilateral ( $n = 5$  birds) Nif lesions. Repeated measures ANOVA revealed that lesions resulted in significant reductions of completion rates in unilateral ( $F_{(8,16)} = 7.0$ ,  $P = 5 \times 10^{-4}$ ), but not bilateral ( $F_{(6,18)} = 4.1$ ,  $P = 0.07$ ), lesions. Post-hoc analysis of the unilateral lesion data using Dunnett's test showed motif completion rates to be significantly different from pre-lesion on the day of lesion ( $P = 5 \times 10^{-4}$ ), but not thereafter ( $P > 0.11$ ). Bottom, comparison of post-lesion motif tempo to pre-lesion baseline for birds with unilateral ( $n = 4$  birds) and bilateral ( $n = 5$  birds) Nif lesions. Repeated measures ANOVA revealed that Nif lesions had a significant effect on motif tempo in both unilateral ( $F_{(8,16)} = 10.4$ ,  $P = 4.6 \times 10^{-5}$ ) and bilateral ( $F_{(6,18)} = 17.5$ ,  $P = 1.3 \times 10^{-6}$ ) conditions. Post-hoc analysis using Dunnett's test showed motif tempo was slowed down for both unilateral and bilateral lesions: the effects remained significant ( $P < 0.05$ )

throughout the 7 days in the bilaterally lesioned birds, and through the first 4 days in unilaterally lesioned birds.

**Code availability.** All custom-written code will be made available upon request.

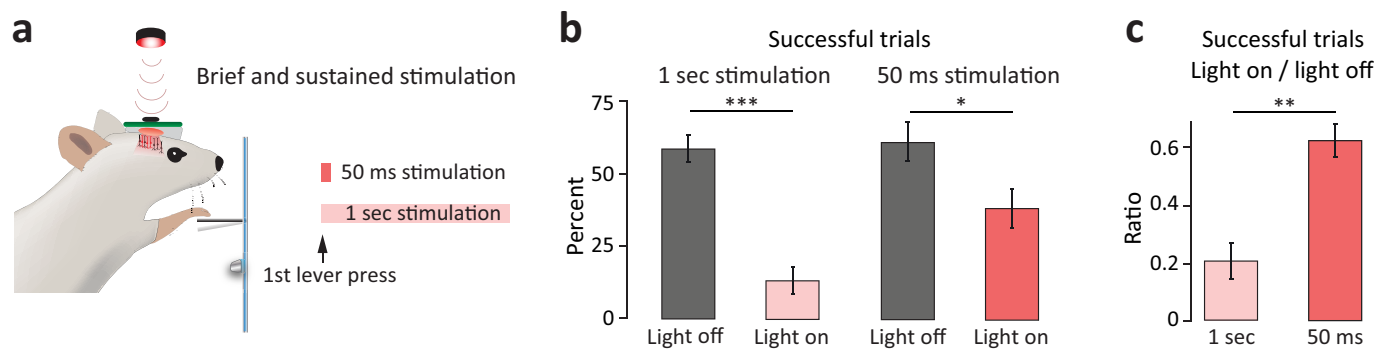
51. Poddar, R., Kawai, R. & Ölveczky, B. P. A fully automated high-throughput training system for rodents. *PLoS ONE* **8**, e83171 (2013).
52. Neafsey, E. J. *et al.* The organization of the rat motor cortex: a microstimulation mapping study. *Brain Res.* **396**, 77–96 (1986).
53. Allen, T. A. *et al.* Imaging the spread of reversible brain inactivations using fluorescent muscimol. *J. Neurosci. Methods* **171**, 30–38 (2008).
54. Ölveczky, B. P., Andalman, A. S. & Fee, M. S. Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLoS Biol.* **3**, e153 (2005).
55. Naie, K. & Hahnloser, R. H. R. Regulation of learned vocal behavior by an auditory motor cortical nucleus in juvenile zebra finches. *J. Neurophysiol.* **106**, 291–300 (2011).
56. Ravbar, P., Lipkind, D., Parra, L. C. & Tchernichovski, O. Vocal exploration is locally regulated during song learning. *J. Neurosci.* **32**, 3422–3432 (2012).
57. Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B. & Mitra, P. P. A procedure for an automated measurement of song similarity. *Anim. Behav.* **59**, 1167–1176 (2000).
58. Fiete, I. R., Senn, W., Wang, C. Z. H. & Hahnloser, R. H. R. Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity. *Neuron* **65**, 563–576 (2010).





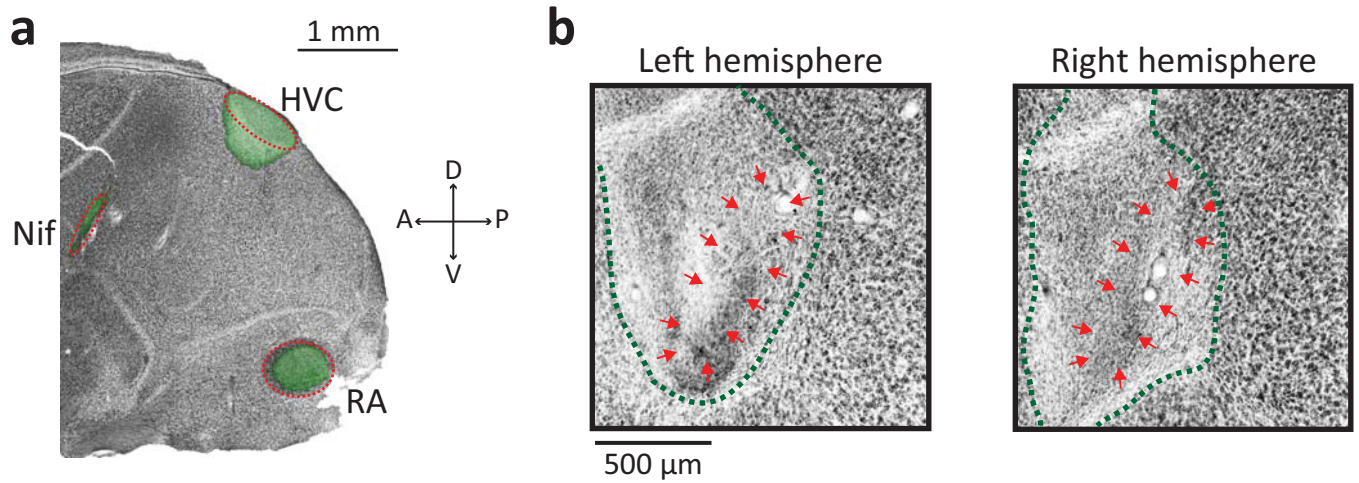
**Extended Data Figure 1 | Light stimulation of motor cortical neurons expressing the optogenetic activator Chrimsom.** **a**, Representative example of AAV-injections into motor cortex, showing Chrimsom-tdTomato expression at different magnifications in a coronal brain section ( $\sim 1.5$  mm anterior to bregma). The scheme of the brain (right) is adapted from Paxinos' rat atlas. The estimated spread of the injections was  $8.3 \pm 1.3$  mm<sup>3</sup> (mean  $\pm$  s.d.,  $n = 2$  rats), with an average of  $31 \pm 2\%$  infected cells (Methods). **b**, Heatmap showing the instantaneous firing rates of

28 single units recorded in an anaesthetized rat in response to a 1 s light pulse, averaged over 30 stimulations (Methods). **c**, A custom-built battery-operated wireless optogenetic stimulation device, consisting of a printed circuit board with integrated IR sensor and LED ( $\lambda = 615$  nm). The IR sensor gates the circuit and allows the LED to be triggered by an IR light-source placed on top of the rat's cage. During surgery, the LED is affixed atop a small craniotomy above motor cortex.



**Extended Data Figure 2 | Both brief and sustained optogenetic stimulation of motor cortex cause significant performance deficits in our task.** **a**, Optogenetic stimulation was triggered on the first lever press in a trial, and lasted for either 50 ms or 1 s. **b**, Both sustained (1 s, left, compare Fig. 1h,  $n=5$  rats,  $P=3 \times 10^{-5}$ , paired  $t$ -test) and brief (50 ms, right,  $n=3$

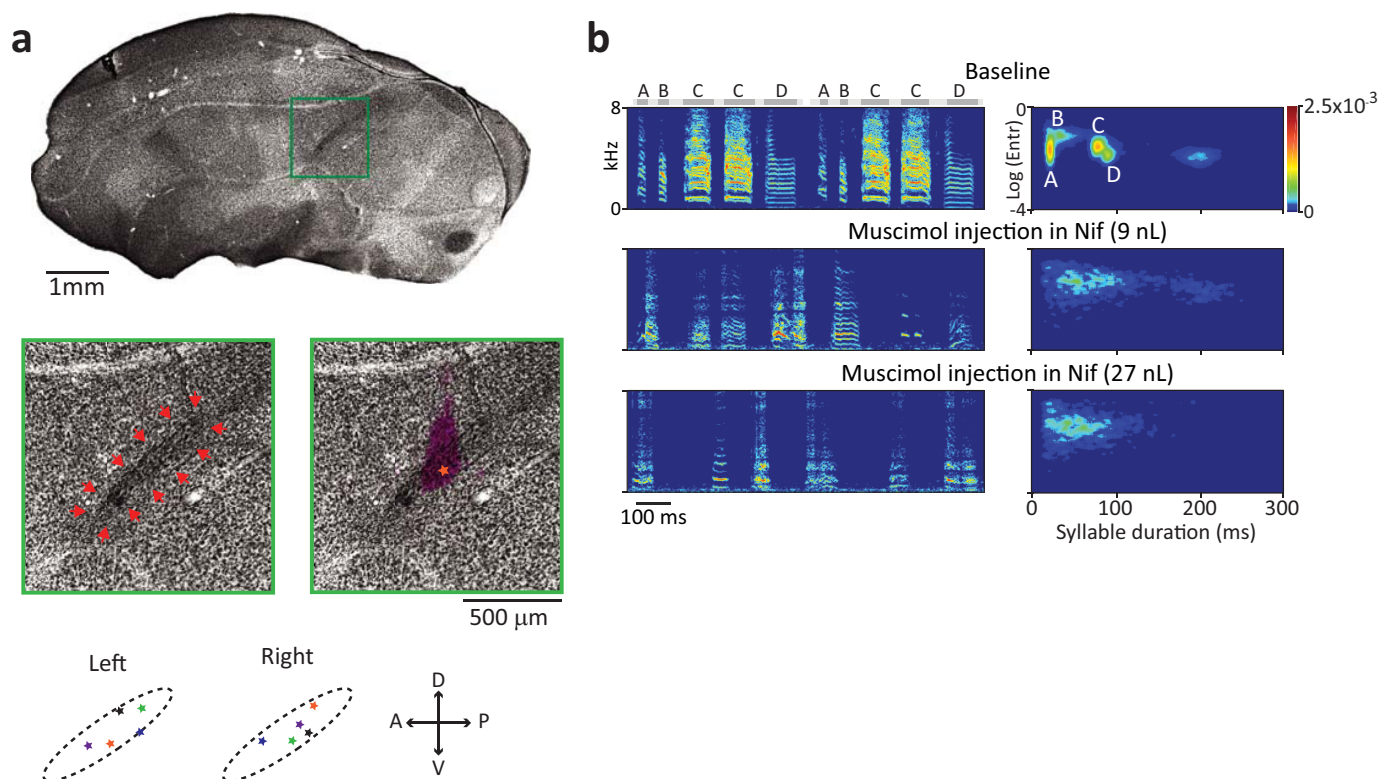
rats,  $P=0.01$ , paired  $t$ -test) optogenetic activation of motor cortex interfered with normal task performance. **c**, Comparing the effects of the two stimulation protocols on task performance (ratio light on/light off) shows that sustained stimulation has a significantly larger effect (1 s:  $n=5$ ; 50 ms:  $n=3$ ,  $P=0.004$ , unpaired  $t$ -test). Error bars represent s.e.m.



**Extended Data Figure 3 | Localization and lesioning of Nif.** **a**, Injection of fluorescently labelled cholera toxin subunit B (green) into HVC retrogradely labels Nif and anterogradely labels downstream control nucleus RA. **b**, Bilateral injections of the excitotoxin NMA produced focal

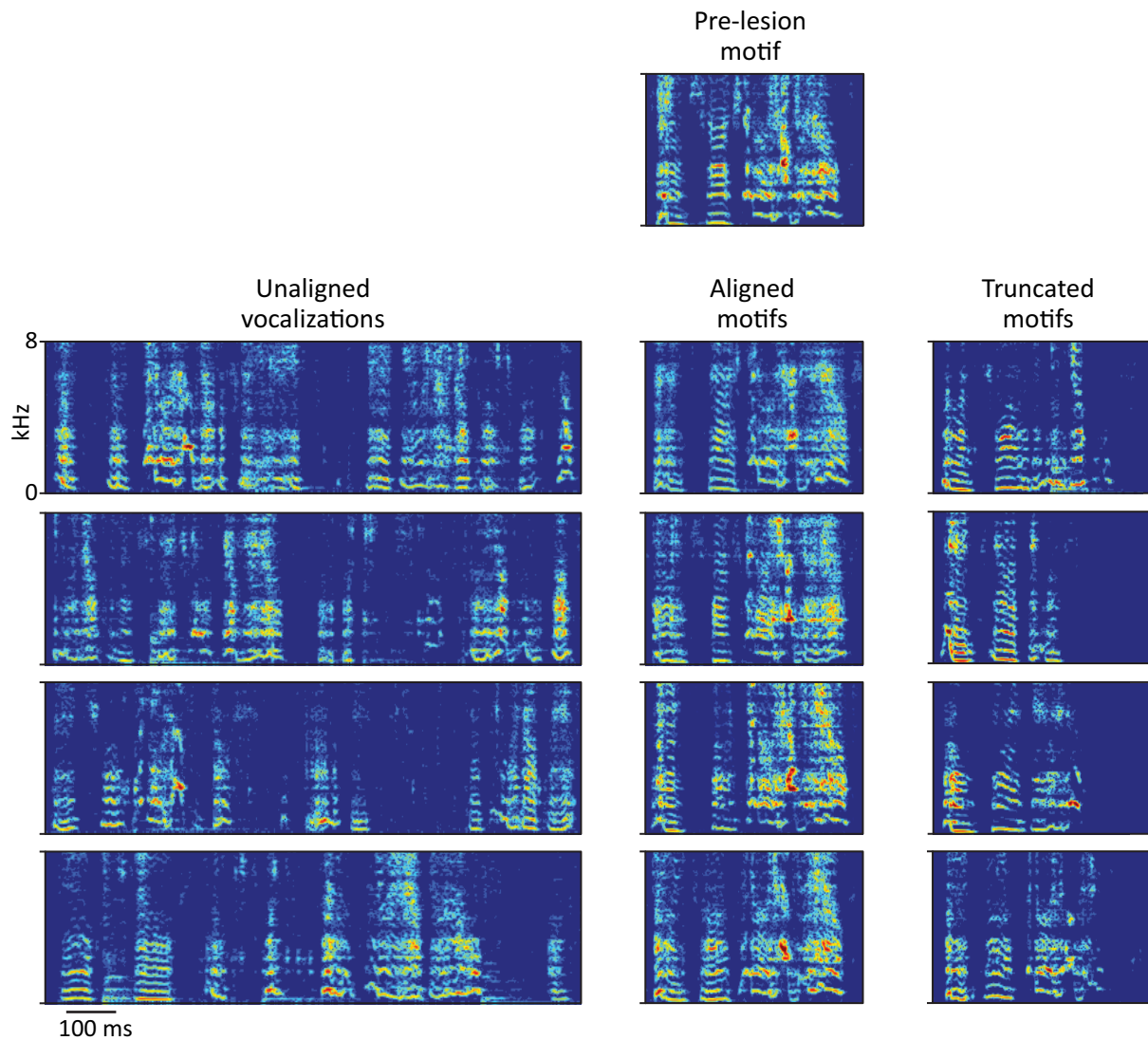
lesions of Nif. Shown are Nissl stained sections from both hemispheres in the same example bird. Red arrows indicate the estimated boundaries of Nif; dashed green line shows the extent of the lesion.





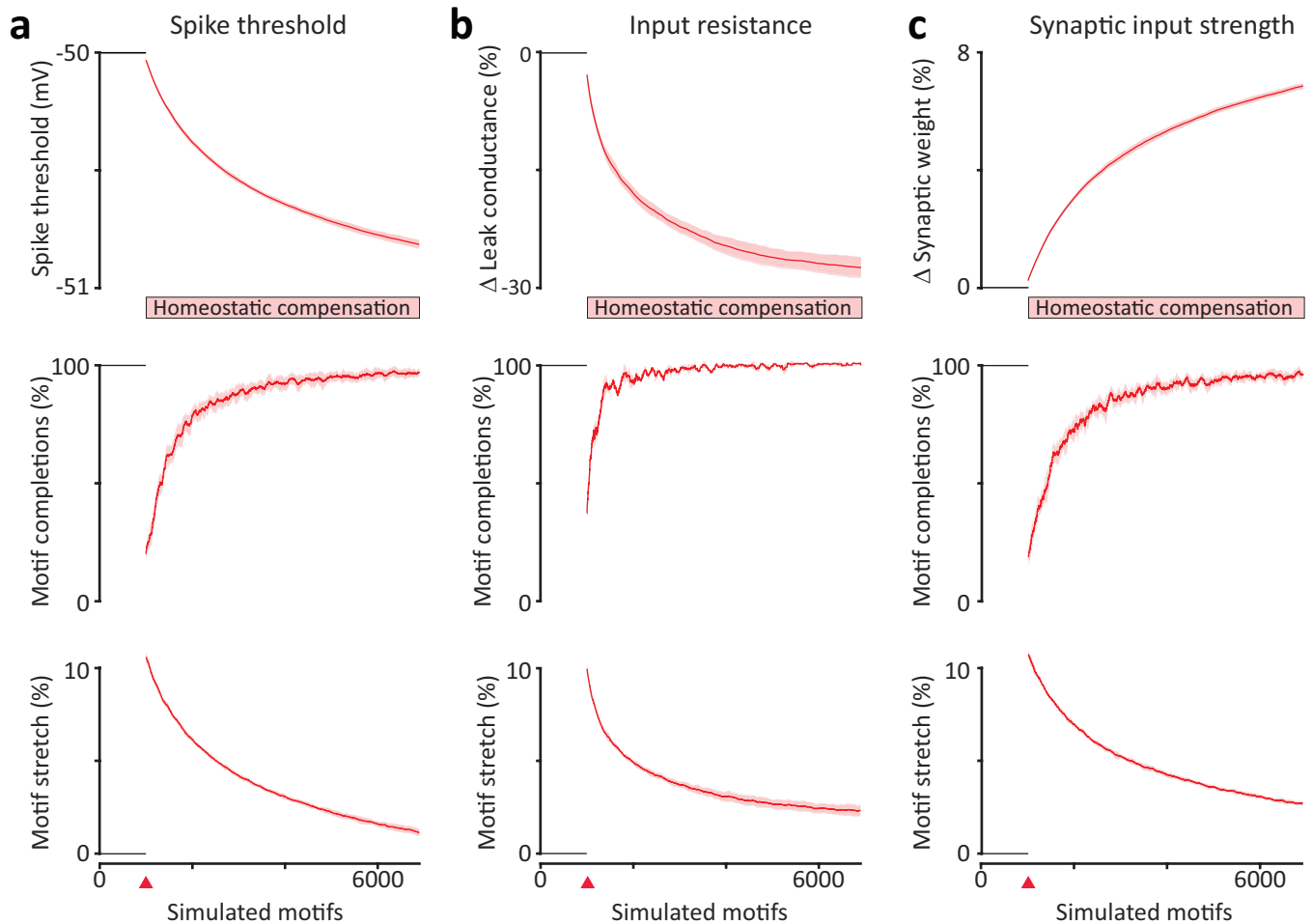
**Extended Data Figure 4 | Muscimol injections into Nif.** **a**, Top, Nissl-stained parasagittal section of a zebra finch brain. Middle, magnified view of the region demarcated with a green square atop. Red arrows (left) indicate the estimated boundaries of Nif; violet overlay (right) shows the spread of fluorescent dye co-injected with muscimol. Orange star indicates estimated centre of injection based on brightness of the fluorescence. Bottom, estimated injection sites relative to the boundaries of Nif for all muscimol-injected birds. Colours denote different animals. **b**, Syllable

spectrograms (left) and entropy-duration distributions (right) for a bird injected with different volumes of muscimol in Nif. Example spectrograms for 9 nl and 27 nl injections are from recordings made 3 min and 7 min after the injections, respectively. That song disruption was similarly rapid and severe for both volumes (in conjunction with the lack of effect from injections above Nif) limits the possibility that the effects on song were due to diffusion of the drug into HVC.



**Extended Data Figure 5 | Spectrograms of vocalizations following unilateral Nif lesion.** Data for the example bird in Fig. 3. All examples were recorded within the first hour of singing after lesion. Top, example spectrogram of a motif recorded just before lesion. Left, example spectrograms of vocalizations in which motif syllables could not be

reliably identified and thus were excluded from subsequent analysis. Middle, example spectrograms of identifiable motifs that were included in the alignment-dependent analysis (Fig. 3f–j). Right, example spectrograms of songs with identifiable syllables, but truncated motifs.



**Extended Data Figure 6 | Different mechanisms for homeostatic regulation of neural activity produce similar effects.** **a**, Top, effect of Nif removal on membrane excitability during simulated songs in a model neuron (from the 40th node), smoothed with a 100-point moving average and averaged over 40 model 'experiments' (shaded regions denote standard deviation across 'experiments'). A rule for homeostatic regulation of activity drives a reduction in spiking threshold after Nif

removal. Middle, fraction of simulations in a 100-point window for which activity in the model HVC network propagated to the end, averaged over 40 model 'experiments' (Methods). Bottom, a 100-point moving average over the time to complete a full chain propagation, averaged over 40 model 'experiments'. Orange triangle denotes time of Nif removal. Same as in Fig. 4b, d. **b**, **c**, Same as in **a**, but with homeostatic regulation of membrane leak conductance (**b**) and synaptic input strength (**c**) (Methods).



# Functional overlap of the *Arabidopsis* leaf and root microbiota

Yang Bai<sup>1\*</sup>, Daniel B. Müller<sup>2\*</sup>, Girish Srinivas<sup>1\*</sup>, Ruben Garrido-Oter<sup>1,3,4\*</sup>, Eva Potthoff<sup>2</sup>, Matthias Rott<sup>1</sup>, Nina Dombrowski<sup>1</sup>, Philipp C. Münch<sup>5,6,7</sup>, Stijn Spaepen<sup>1</sup>, Mitja Remus-Emsermann<sup>2</sup>, Bruno Hüttel<sup>8</sup>, Alice C. McHardy<sup>4,5</sup>, Julia A. Vorholt<sup>2\*</sup> & Paul Schulze-Lefert<sup>1,4\*</sup>

**Roots and leaves of healthy plants host taxonomically structured bacterial assemblies, and members of these communities contribute to plant growth and health. We established *Arabidopsis* leaf- and root-derived microbiota culture collections representing the majority of bacterial species that are reproducibly detectable by culture-independent community sequencing. We found an extensive taxonomic overlap between the leaf and root microbiota. Genome drafts of 400 isolates revealed a large overlap of genome-encoded functional capabilities between leaf- and root-derived bacteria with few significant differences at the level of individual functional categories. Using defined bacterial communities and a gnotobiotic *Arabidopsis* plant system we show that the isolates form assemblies resembling natural microbiota on their cognate host organs, but are also capable of ectopic leaf or root colonization. While this raises the possibility of reciprocal relocation between root and leaf microbiota members, genome information and recolonization experiments also provide evidence for microbiota specialization to their respective niche.**

Plants and animals harbour abundant and diverse bacterial microbiota<sup>1</sup>. These taxonomically structured bacterial communities have important functions for the health of their multicellular eukaryotic hosts<sup>2–4</sup>. The leaf and root microbiota of flowering plants have been extensively studied by culture-independent analyses, which have consistently revealed the co-occurrence of four main bacterial phyla: Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria<sup>5–15</sup>. Determinants of microbiota composition at lower taxonomic ranks, that is, at genus and species level, are host compartment, environmental factors and host genotype<sup>6,7,12,16</sup>.

Soil harbours an extraordinary rich diversity of bacteria and these define the start inoculum of the *Arabidopsis thaliana* root microbiota<sup>6,7</sup>. The inoculum source of the leaf microbiota is thought to be more variable owing to the inherently open nature of the leaf ecosystem, probably involving bacteria transmitted by aerosols, insects, or soil<sup>8,9,17</sup>. A recent study of the grapevine (*Vitis vinifera*) microbiota showed that the root-associated bacterial assemblies differed significantly from aboveground communities, but that microbiota of leaves, flowers, and grapes shared a greater proportion of taxa with soil communities than with each other, suggesting that soil may serve as a common bacterial reservoir for belowground and aboveground plant microbiota<sup>18</sup>.

A major limitation of current plant microbiota research is the lack of systematic microbiota culture collections that can be employed in microbiota reconstitution experiments with germ-free plants to address principles underlying community assembly and proposed microbiota functions for plant health under laboratory conditions<sup>19</sup>.

## Bacterial culture collections from roots and leaves

We employed three bacterial isolation procedures to establish taxonomically diverse culture collections of the *A. thaliana* root and leaf microbiota. Bacterial isolates were recovered from pooled or individual

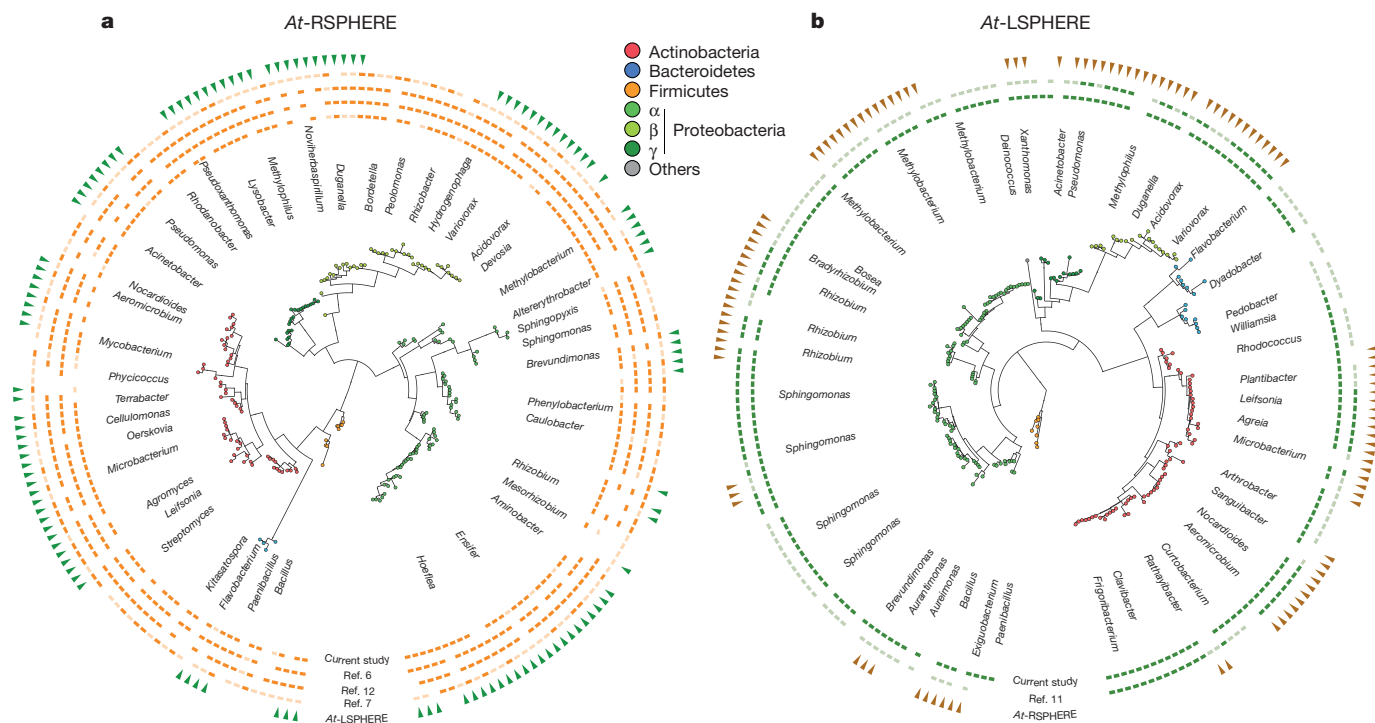
roots or leaves of healthy plants using colony picking from agar plates, limiting dilution in liquid media in 96-well microtitre plates, or microbial cell sorting (see Methods). We adopted a two-step bar-coded pyrosequencing protocol<sup>20</sup> for taxonomic classification of the cultured bacteria by determining  $\geq 550$  base pairs (bp) 16S ribosomal RNA (rRNA) gene sequences (Supplementary Fig. 1; Methods). In parallel, parts of the root and leaf material was used for cultivation-independent 16S rRNA gene community sequencing to cross-reference Operational Taxonomic Unit (OTU)-defined taxa from the microbiota with individual colony forming units (CFUs) in the culture collections.

A total of 5,812 CFUs were recovered from 59 independently pooled *A. thaliana* root samples of plants mainly grown in Cologne soil, Germany, whereas 2,131 CFUs were retrieved from leaf washes of individual leaves collected from *A. thaliana* populations at six locations near Tübingen, Germany, or Zurich, Switzerland (Supplementary Data 1). Recovery estimates for root-associated OTUs were calculated using the culture-independent community profiles of the present and two earlier studies<sup>6,12</sup> and varied for the top 100 OTUs (70% of sequencing reads) between 54–65% and at  $\geq 0.1\%$  relative abundance (RA) between 52–64% (Methods; Extended Data Fig. 1a–c; Supplementary Data 2). For leaf samples, the culture-independent 16S rRNA gene analyses from individual and pooled leaves (60 samples from six sites) revealed similar community profiles at all tested geographic sites and high leaf-to-leaf consistency (Extended Data Fig. 2). Recovery estimates of the top 100 leaf-associated bacterial OTUs (86% of all sequencing reads) were 54% and at  $\geq 0.1\%$  RA 47% (Extended Data Fig. 1d). The root-derived CFUs correspond to 23 of 38 and the leaf-derived CFUs belong to 28 of 45 detectable bacterial families. Root- and leaf-derived CFUs each represent all four bacterial phyla typically associated with *A. thaliana* roots and leaves. Thus, most bacterial families that are reproducibly associated with *A. thaliana* roots and leaves have culturable members.

<sup>1</sup>Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. <sup>2</sup>Institute of Microbiology, ETH Zurich, 8093 Zurich, Switzerland.

<sup>3</sup>Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. <sup>4</sup>Cluster of Excellence on Plant Sciences (CEPLAS), Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. <sup>5</sup>Computational Biology of Infection Research, Helmholtz Center for Infection Research, 38124 Braunschweig, Germany. <sup>6</sup>Max-von-Pettenkofer Institute, Ludwig Maximilian University, German Center for Infection Research (DZIF), partner site LMU Munich, 80336 Munich, Germany. <sup>7</sup>German Center for Infection Research (DZIF), partner site Hannover-Braunschweig, 38124 Braunschweig, Germany. <sup>8</sup>Max Planck Genome Center, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany.

\*These authors contributed equally to this work.



**Figure 1 | Taxonomic overlap between *At*-RSPHERE and *At*-LSPHERE isolates and their representation in culture-independent microbiota profiling studies. a, b, Phylogenetic trees of *At*-RSPHERE (a;  $n = 206$  isolates) and *At*-LSPHERE (b;  $n = 224$  isolates) bacteria. Their taxonomic overlap is shown in the outermost ring (green or brown triangles). a, Representation of *At*-RSPHERE bacteria in each of four indicated culture-independent profiling studies of the *A. thaliana* root microbiota;**

### *At*-RSPHERE and *At*-LSPHERE culture collections

We selected from the aforementioned culture collections a taxonomically representative core set of bacterial strains after Sanger sequencing of a  $\geq 550$  bp fragment of the 16S rRNA gene and additional strain purification (Methods). To increase the intra-species genetic diversity of the culture collections, and because the quantitative contribution of a single isolate to its corresponding OTU cannot be estimated, we included bacterial strains sharing  $\geq 97\%$  16S rRNA gene sequence identity (widely used for bacterial species definition), but representing independent host colonization events, that is, recovered from different plant roots or leaves. In total we selected 206 root-derived isolates that comprise 28 bacterial families belonging to four phyla (designated *At*-RSPHERE) and 224 leaf-derived isolates that comprise 29 bacterial families belonging to five phyla (designated *At*-LSPHERE) (Extended Data Fig. 3a, b; Supplementary Data 1; Methods). Additionally, to represent abundant soil OTUs ( $\geq 0.1\%$  RA) we selected 33 bacterial isolates encompassing eight bacterial families belonging to three phyla from unplanted Cologne soil (Extended Data Fig. 3c).

Notably, the majority of the *At*-RSPHERE isolates share  $\geq 97\%$  16S rRNA gene sequence identity matches with root-associated OTUs reported in four independent studies in which *A. thaliana* plants had been grown in Cologne soil<sup>6,12</sup> or other European<sup>6,12</sup> or US soils<sup>7</sup> (inner four circles in Fig. 1a; Methods). Similarly, the bulk of *At*-LSPHERE isolates match leaf-derived OTUs detected in *A. thaliana* populations at the Tübingen/Zurich locations or US-grown plants (innermost two circles in Fig. 1b). This indicates that representatives of the majority of *At*-RSPHERE and *At*-LSPHERE members co-populate the corresponding *A. thaliana* organs in multiple tested environments, including two continents, Europe and North America.

Phylogenetic analysis based on 16S rRNA gene Sanger sequences revealed that 119 out of 206 *At*-RSPHERE isolates (58%) share  $\geq 97\%$  sequence identity matches with corresponding 16S rRNA gene

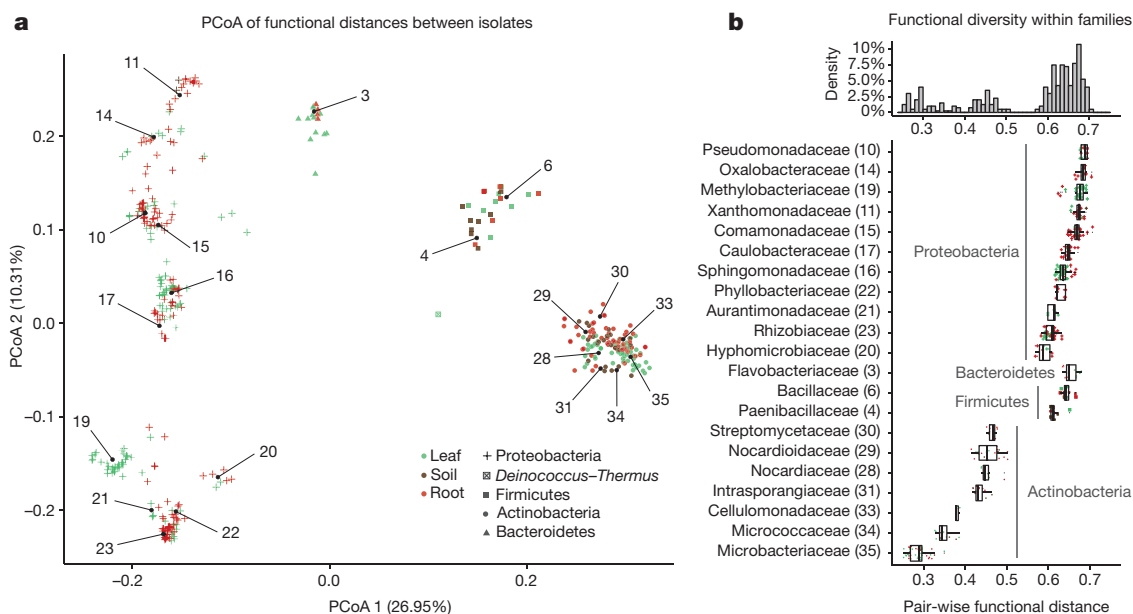
root-associated OTUs with RAs  $\geq 0.1\%$  (dark orange) or  $\leq 0.1\%$  (light orange). b, Representation of *At*-LSPHERE bacteria in the two indicated culture-independent phyllosphere profiling studies; leaf-associated OTUs with RAs  $\geq 0.1\%$  (dark green) or  $< 0.1\%$  (light green). Taxonomic assignment and phylogenetic tree inference were based on partial 16S rRNA gene Sanger sequences.

fragments of *At*-LSPHERE members (outermost circle in Fig. 1a). Similarly, 108 out of 224 *At*-LSPHERE isolates (48%) share  $\geq 97\%$  sequence identity matches with *At*-RSPHERE members (outermost circle in Fig. 1b). This extensive overlap both at the rank of bacterial genera and bacterial families (20 out of 38 detectable families) between leaf- and root-derived bacteria is notable because we collected leaf and root specimen from environments that are geographically widely separated ( $> 500$  km) and is consistent with a previous report on leaf and root microbiota overlap in *V. vinifera*<sup>18</sup>. This overlap is corroborated by the corresponding culture-independent leaf and root community profiles (Extended Data Fig. 4). As essentially all *A. thaliana* root-associated bacteria are recruited from the surrounding soil biome<sup>6,7,12</sup>, this raises the possibility that unplanted soil also defines the start inoculum for a substantial proportion of the leaf microbiota with subsequent selection for niche-adapted organisms.

### Comparative genome analysis of the culture collections

To characterize the functional capabilities of the core culture collections we subjected each isolate to whole-genome sequencing and generated a total of 432 high-quality draft genomes (206 from leaf, 194 from root and 32 from soil; Supplementary Data 3). Taxonomic assignment of the whole-genome sequences confirmed that these isolates span a broad taxonomic range, belonging to 35 different bacterial families distributed across five phyla (Supplementary Data 4).

Based on the whole-genome taxonomic information, we grouped the isolates into family-level clusters. We found that clusters of genomes are characterized by a relatively large core-genome, with an average of 33.6% of the annotated proteins present in each member and a smaller fraction of singleton genes identified in only one genome per cluster (14.0%). Detailed analysis of phylogenetic diversity of each cluster revealed a substantial overlap between leaf, root and soil isolates (Supplementary Data 5). Many clusters showed no clear separation of



**Figure 2 | Analysis of functional diversity between sequenced isolates.** **a**, Principal coordinate analysis (PCoA) plot depicting functional distances between sequenced genomes ( $n = 432$ ) based on the KEGG Orthology (KO) database annotation. Each point represents a genome. Colours represent the organ of isolation and shapes correspond to their taxonomy. Numbers inside the plot refer to bacterial families listed in **b**. **b**, Analysis of functional diversity within bacterial families as measured by pair-wise

isolates based on their ecological niche, suggesting shared core functions. However, other clusters contained isolates of one organ or showed clear separation among them, suggesting niche specialization within some clusters (Supplementary Data 5). We then examined the functional diversity between the sequenced isolates in order to determine whether the observed phylogenetic overlap corresponded with functional similarities between leaf and root isolates. Principal coordinates analysis (PCoA) of functional distances (Fig. 2a; Methods) revealed a clear clustering of genomes on the basis of their taxonomy, but only limited separation of genomes on the basis of their ecological compartment. Taken together, both phylogenetic and functional diversification of the genomes is strongly driven by their taxonomic affiliation and weakly by the ecological niche.

We examined the functional diversity within each bacterial family (Fig. 2b) in order to identify bacterial taxa with varying degrees of functional versatility. Families belonging to Actinobacteria show a lower functional diversity (average distance 0.37) compared to those belonging to Bacteroidetes, Firmicutes and especially Proteobacteria (0.65 average pair-wise distance), which exhibit a higher degree of within-family functional diversification, even though all family-level groups have a comparable degree of phylogenetic relatedness. Among these groups, Pseudomonadaceae, Oxalobacteraceae and Methylobacteriaceae members show the highest functional heterogeneity, compared to Microbacteriaceae strains, which we identified as the least functionally diverse family (Fig. 2b).

We searched for signatures of niche specialization at individual functional categories using enrichment analysis to identify functional categories over-represented in genomes from root and leaf or soil isolates (Fig. 3; Methods). Specifically, we found the category 'carbohydrate metabolism' to be enriched in the leaf and soil genomes compared to those isolated from roots (Mann-Whitney test,  $P = 1.29 \times 10^{-7}$ ; Fig. 3b). We speculate that this differential enrichment could reflect the availability of simple carbon sources in roots through the process of root exudation (sugars, amino acids, aliphatic acids)<sup>21,22</sup>, whereas bacteria associated with leaves or unplanted soil might rely on a more diverse repertoire of carbohydrate metabolism genes to access scarce

functional distances between genomes (bottom panel;  $n = 432$ ). Higher pairwise distances between members of a family indicate a larger degree of functional diversity. Only families with at least five members are shown. The histogram (top panel) was calculated for the entire data set and the y-axis corresponds to the percentage of data points in each bin. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.

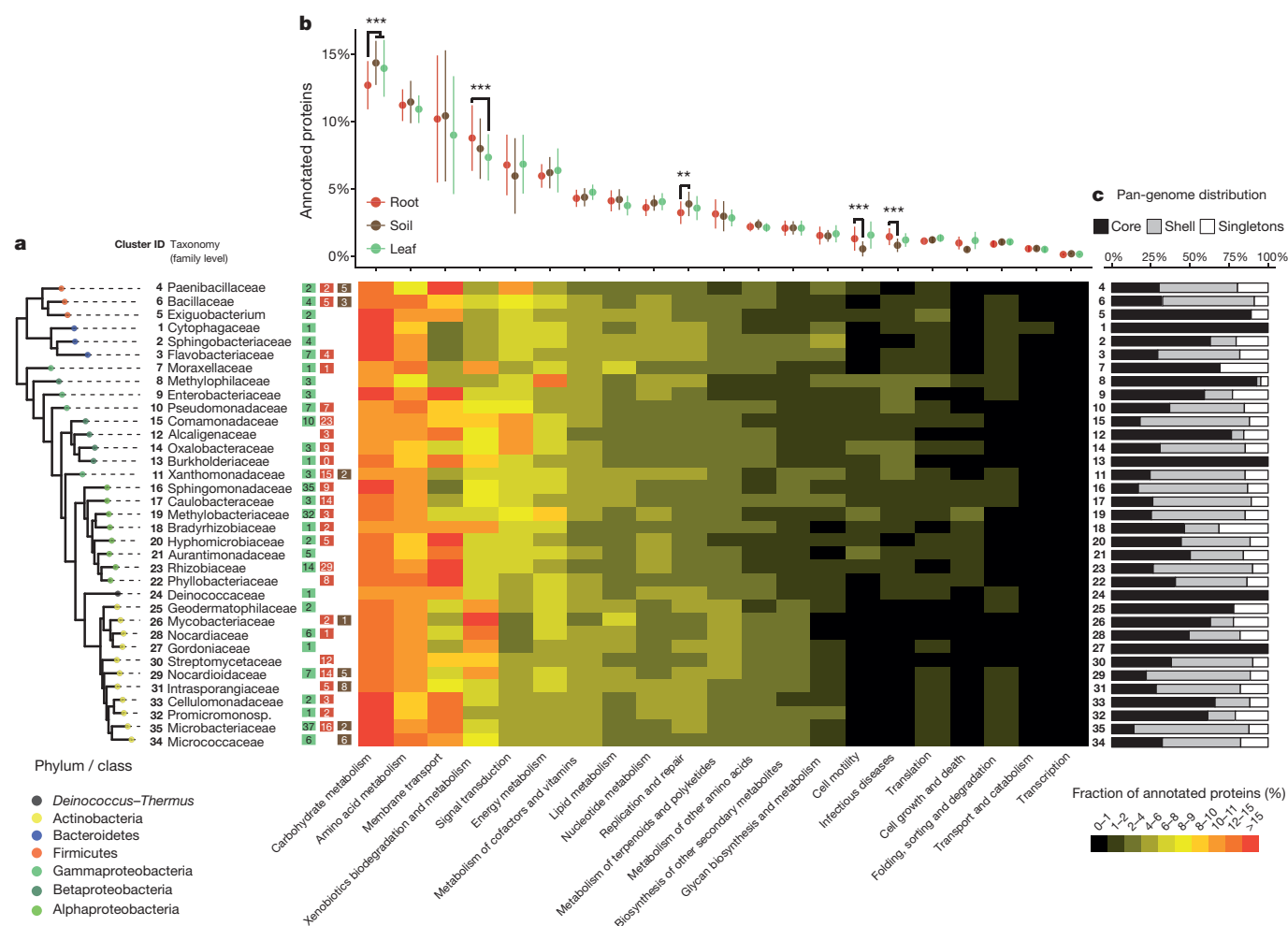
and complex organic carbon, for example, polysaccharides and leaf cuticular waxes. The category 'xenobiotics biodegradation and catabolism' is enriched in the root genomes with respect to those isolated from leaves ( $P = 2.60 \times 10^{-11}$ ; Fig. 3b), which is consistent with previous evidence that genes for aromatic compound utilization are expressed in the rhizosphere<sup>23</sup>. No single taxon is responsible for these significant differences, but this seems to be a general feature across the sequenced bacterial genomes of the respective ecological niche (Extended Data Figs 5 and 6). Interestingly, we observed the same trends of differential abundance of functional categories in *V. vinifera* root metagenome samples<sup>18</sup> compared to their respective unplanted soil controls (Extended Data Fig. 7).

Together, these findings indicate a substantial overlap of functional capabilities in the genomes of the *Arabidopsis* leaf- and root-derived culture collections and differences at the level of individual functional categories that may reflect specialization of the leaf and root microbiota to their respective niche. Additional genomic signatures for niche-specific colonization are likely to be hidden in genes for which a functional annotation is currently unavailable (~57%).

### Synthetic community colonization of germ-free plants

We colonized germ-free *A. thaliana* plants with synthetic communities (SynComs) consisting of bacterial isolates from our culture collections to assess their potential for host colonization in a gnotobiotic system containing calcined clay as inert soil substitute (Methods). To mimic the taxonomic diversity of leaf and root microbiota in natural environments we employed mainly two SynComs: 'L' comprising 218 leaf-derived bacteria and 'R+S' consisting of 188 members of which 158 are root-derived and 30 are soil-derived bacteria (Supplementary Data 6). Input SynComs were either inoculated directly before sowing of surface-sterilized seeds in calcined clay and/or spray-inoculated on leaves of three-week-old germ-free plants. For all defined communities we examined three independent SynCom preparations, each tested in three closed containers containing four plants. We employed 16S rRNA gene community profiling with a method validated for defined communities<sup>24</sup> to detect potential community shifts between input and output





**Figure 3 | Functional analysis of sequenced isolates. a**, Phylogeny of family-level clusters of bacterial isolates. The tips of the tree are annotated, from left to right, with the cluster ID, taxonomic classification, followed by the number of sequenced isolates from leaf, root or soil that constitute each cluster. The heat map depicts the average percentage of annotated proteins of each cluster belonging to each functional category. **b**, Functional enrichment analysis between leaf ( $n = 206$ ), root ( $n = 194$ )

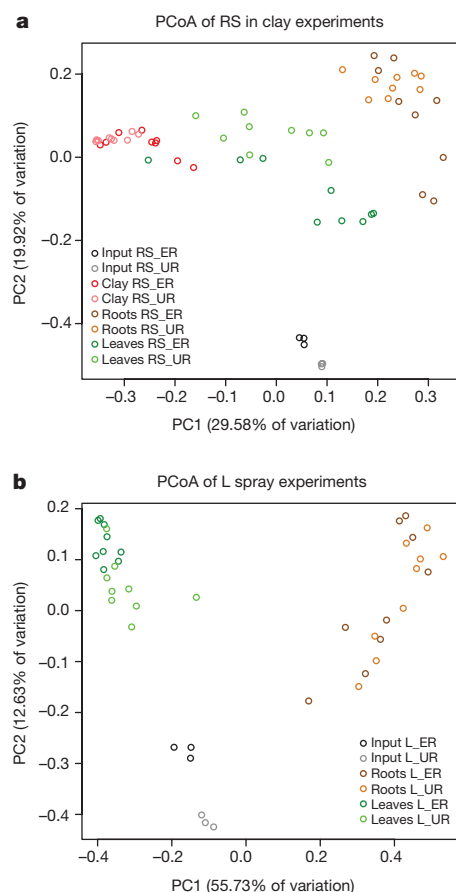
and soil ( $n = 32$ ) genomes. Points and bars correspond to the mean abundance and standard deviation of each functional category.  $P$  values were obtained using the non-parametric Mann–Whitney test corrected by the Bonferroni approach. **c**, Analysis of pan-genome distribution for each cluster of genomes, indicating the percentage of annotated proteins found in only one isolate (singletons), in more than one but not all (shell) or in all genomes within the cluster (core).

SynComs in samples of seven week-old roots, leaves, or unplanted clay. In this community analysis, ‘indicator OTUs’ either represent a single strain or a known group of isolates (Supplementary Data 6).

Upon application of the input R+S SynCom to clay (‘R+S in clay’) and co-cultivation with *A. thaliana* plants for seven weeks we retrieved reproducible R+S output communities from clay (without host), root, and leaf compartments (Supplementary Fig. 2). These output SynCom profiles were robust against a 75% reduction in RA of Proteobacteria compared to Actinobacteria, Bacteroidetes and Firmicutes in the input R+S SynCom (input ratios 1:1:1 or 1:1:1:0.25, respectively), which was confirmed by PCoA (Fig. 4a). PCoA also revealed distinct output communities in each of the three tested compartments (Fig. 4a;  $P < 0.001$  Extended Data Fig. 8a, b). This indicates that a marked host-independent community change occurred in clay (without host) as well as host-dependent community shifts that are specific for leaves and roots. Next, we tested the ‘L SynCom of leaf-derived bacteria by spray inoculation on leaves of three week-old plants. After four weeks of L SynCom co-incubation with plants, output communities were detected in leaves and roots (Supplementary Fig. 3). PCoA revealed that these two output communities were different between each other, but robust against a 75% reduction in RA of input Proteobacteria (Fig. 4b; Supplementary Fig. 3;  $P < 0.001$ ;

Extended Data Fig. 8c, d). The converging output communities despite varying RAs of input SynComs suggest that the communities have reached a steady state. These experiments also reveal that both R+S and L SynCom members not only colonize cognate host organs, but are capable of ectopic colonization of leaves and roots, which might be linked to the extensive species overlap of *A. thaliana* leaf and root microbiota in natural environments (Fig. 1a, b). Additionally, this provides experimental support for the hypothesis that a subset of leaf-colonizing bacteria originates from unplanted soil and raises the possibility for reciprocal bacterial colonization events between roots and leaves during and/or after the establishment of the respective microbiota, for example, by ascending migration of rhizobacteria from roots to leaves<sup>25</sup>. Upon leaf spray application of SynComs, a small amount of leaf bacteria is likely to land on the clay surface and thereafter colonize roots, which is not fundamentally different from processes occurring in natural environments, for example, during rain showers and/or leaf dehiscence.

A comparison of rank abundance profiles between indicator OTUs for all root- and leaf-derived isolates and corresponding OTUs identified in the environmental root and leaf samples revealed similar trends at phylum, class and family levels (Extended Data Fig. 9). This validates the gnotobiotic plant system as a tool for microbiota reconstitution experiments.



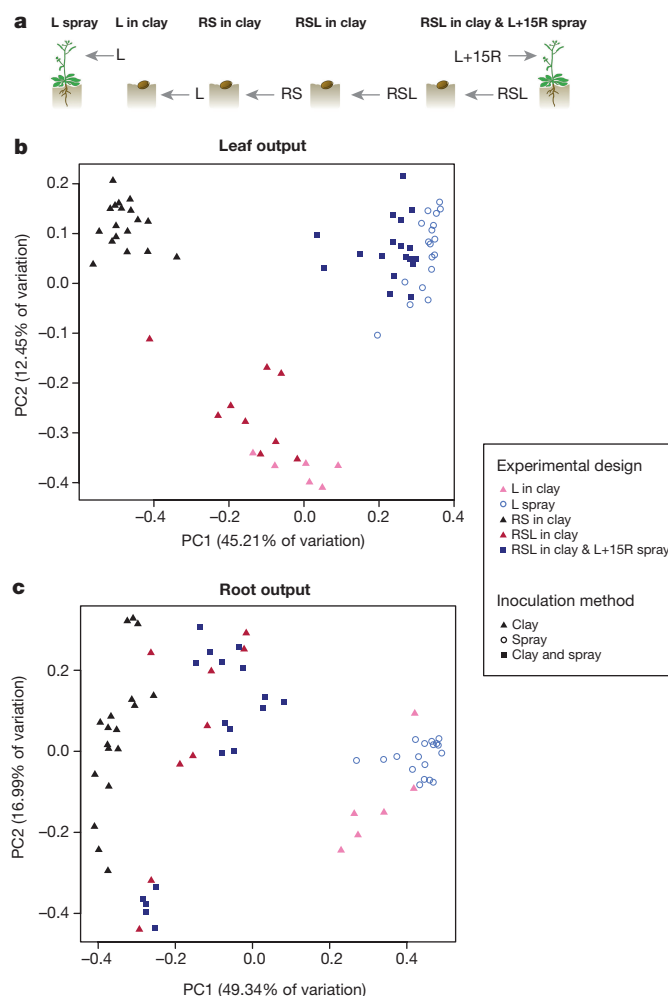
**Figure 4 | SynCom colonization of germ-free *A. thaliana* plants.**

**a, b,** Principal coordinate analysis (PCoA) of Bray–Curtis distances of input and output SynCom profiles of RS in clay (**a**;  $n = 60$ ) and L spray (**b**;  $n = 42$ ) experiments. Each condition was tested with 6 independently prepared SynComs; each preparation was used for 3 independent inoculations. L, leaf-derived strains; RS, root- and soil-derived strains; ER, equal strain ratio; UR, unequal strain ratio.

### Niche-specific microbiota establishment with SynComs

The species overlap between root and leaf microbiota and their corresponding culture collections (Fig. 1a, b; Extended Data Fig. 4) prompted us to test whether R+S and L SynComs equally contribute to root and leaf microbiota establishment. Both SynComs were pooled and inoculated in clay together with surface-sterilized *A. thaliana* seeds (designated ‘RSL in clay’, Fig. 5a). We also tested whether a preformed root-associated community can interfere with leaf-associated community establishment. After three weeks of co-cultivation, half of the plants grown with the ‘RSL in clay’ SynCom were treated by leaf-spray inoculation with the L SynCom supplemented with 15 root-derived strains (designated ‘RSL in clay & L+15R spray’). Plant organ-specific output communities were determined after a further four weeks of co-incubation. We also inoculated the L SynCom alone in clay and determined output SynComs (designated ‘L in clay’, Fig. 5a).

We found significant differences between leaf-associated output communities of the ‘RSL in clay’ and ‘RS in clay’ experiments (Fig. 5b;  $P < 0.001$ , Extended Data Fig. 8f; Supplementary Figs 2 and 4) and that the output community on leaves after ‘L in clay’ inoculation is similar to the leaf outputs of ‘RSL in clay’ inoculation (Fig. 5b;  $P < 0.001$ , Extended Data Fig. 8f; Supplementary Figs 4 and 5), indicating that in this comparison the leaf-derived SynCom has a stronger influence on leaf microbiota structure than root- and soil-derived bacteria. However, both ‘RSL in clay’ and ‘L in clay’ leaf outputs are significantly different from the leaf output of the ‘L spray’ experiment (Fig. 5b;  $P < 0.001$ , Extended Data Fig. 8e; Supplementary Figs 3–5), showing that many leaf-derived isolates do not successfully colonize leaves when only



**Figure 5 | SynCom competition supports host-organ-specific community assemblies.**

**a,** Pictograms illustrating ‘L spray’, ‘L in clay’, ‘RS in clay’, ‘RSL in clay’, and ‘RSL in clay & L+15R spray’ SynCom experiments. **b, c,** PCoA of Bray–Curtis distances of leaf (**b**;  $n = 69$ ) and root (**c**;  $n = 69$ ) outputs of the five experiments illustrated in **a**. R, root-derived isolates; S, soil-derived isolates; L, leaf-derived isolates. L in clay was tested with 6 independently prepared SynComs; RSL in clay experiment was tested with 3 independently prepared SynComs, each used for 3 independent inoculations. All other experiments were tested with 6 independently prepared SynComs and each preparation was used for 3 independent inoculations.

inoculated in the clay environment. For example, of the top 16 genera a total of three are grossly underrepresented in leaf outputs of the ‘RSL in clay’ compared to the ‘RSL in clay & L+15R spray’ experiment (*Chryseobacterium*, *Sphingomonas* and *Variovorax*; Supplementary Fig. 6) and these three genera are abundant in the natural leaf microbiota (Extended Data Fig. 4). Finally, leaf outputs were strikingly similar between ‘RSL in clay & L+15R spray’ and ‘L spray’ only experiments (Fig. 5b; Supplementary Figs 3 and 7), indicating that the L+15R SynCom, leaf spray-inoculated three weeks after RSL application to clay, can displace the RSL leaf output. Collectively, these results support the hypothesis that leaf microbiota establishment benefits from air- and soil-borne inoculations<sup>8,17</sup>, although we note that our single application of bacteria to leaves does not mimic the continuous exposure of plant leaves to airborne microorganisms in nature.

A comparison of the root-associated community outputs of the experiments described above revealed that the ‘RSL in clay’ experiment is more similar to root outputs of the ‘RS in clay’ than ‘L in clay’ experiments (Fig. 5c;  $P < 0.001$  Extended Data Fig. 8g), suggesting that the root- and soil-derived SynCom has a stronger influence on root

microbiota structure than the leaf-derived SynCom. In this experiment the fractional contribution of root-specific indicator OTUs increases in the output, but decreases for leaf-specific indicator OTUs, relative to their input, pointing to a potential adaptation of root-derived bacteria for root colonization (Extended Data Fig. 10a; Mann–Whitney;  $P < 0.05$ ). This is further supported by the observation that in the ‘RSL in clay’ experiment root colonization rates for root-specific indicator OTUs are higher compared to those specific for leaves when applying a 0.1% relative abundance threshold in at least one biological replicate (69% and 33%, respectively). Taken together, this suggests that root-derived bacteria are better adapted to colonize their cognate host niche than leaf-derived bacteria. Further comparisons of the root-associated output communities of the ‘L in clay’ and ‘L spray’ experiments (Fig. 5c; Supplementary Figs 3 and 5) revealed similar community composition, indicating convergence of ectopic root-associated community outputs despite different inoculation time points or sites of application. Additional reciprocal transplantation experiments using a ‘R’ (root strains only) SynCom either applied to clay (‘R in clay’) or by spray inoculation (‘R spray’) confirmed the convergence of ectopic community outputs also for root-derived bacteria on leaves (Extended Data Fig. 10 b, c; Supplementary Figs 8 and 9). Convergence of ectopic SynCom outputs is consistent with the hypothesis that a subset of leaf and root colonizing bacteria has the potential to relocate between leaves and roots.

## Conclusions

By employing systematic bacterial isolation approaches, we established expandable culture collections of the *A. thaliana* leaf- and root-associated microbiota, which capture the majority of the species found reproducibly in their respective natural communities ( $\geq 0.1\%$  relative abundance). The sequenced bacterial genomes as well as any future updates are available at <http://www.at-sphere.com>. These resources together with the remarkable reproducibility of the gnotobiotic reconstitution system enable future studies on bacterial community establishment and functions under laboratory conditions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 June; accepted 9 November 2015.

Published online 2 December 2015.

- Rosenberg, E. & Xilber-Rosenberg, I. *The Hologenome Concept: Human, Animal and Plant Microbiota* (Springer, 2013).
- Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Rev. Microbiol.* **9**, 279–290 (2011).
- Berendsen, R. L., Pieterse, C. M. & Bakker, P. A. The rhizosphere microbiome and plant health. *Trends Plant Sci.* **17**, 478–486 (2012).
- Subramanian, S. *et al.* Cultivating healthy growth and nutrition through the gut microbiota. *Cell* **161**, 36–48 (2015).
- Delmotte, N. *et al.* Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc. Natl Acad. Sci. USA* **106**, 16428–16433 (2009).
- Bulgarelli, D. *et al.* Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* **488**, 91–95 (2012).
- Lundberg, D. S. *et al.* Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488**, 86–90 (2012).
- Vorholt, J. A. Microbial life in the phyllosphere. *Nature Rev. Microbiol.* **10**, 828–840 (2012).
- Bodenhausen, N., Horton, M. W. & Bergelson, J. Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. *PLoS One* **8**, e56329 (2013).
- Guttman, D. S., McHardy, A. C. & Schulze-Lefert, P. Microbial genome-enabled insights into plant-microorganism interactions. *Nature Rev. Genet.* **15**, 797–813 (2014).
- Horton, M. W. *et al.* Genome-wide association study of *Arabidopsis thaliana* leaf microbial community. *Nat. Commun.* **5**, 5320 (2014).

- Schlaeppli, K., Dombrowski, N., Oter, R. G., Ver Loren van Themaat, E. & Schulze-Lefert, P. Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives. *Proc. Natl Acad. Sci. USA* **111**, 585–592 (2014).
- Edwards, J. *et al.* Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc. Natl Acad. Sci. USA* **112**, E911–E920 (2015).
- Hacquard, S. *et al.* Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe* **17**, 603–616 (2015).
- Bulgarelli, D. *et al.* Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe* **17**, 392–403 (2015).
- Lebeis, S. L. *et al.* Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science* **349**, 860–864 (2015).
- Maignien, L., DeForce, E. A., Chafee, M. E., Eren, A. M. & Simmons, S. L. Ecological succession and stochastic variation in the assembly of *Arabidopsis thaliana* phyllosphere communities. *MBio* **5**, e00682–e13 (2014).
- Zarraoaindia, I. *et al.* The soil microbiome influences grapevine-associated microbiota. *MBio* **6**, e02527–14 (2015).
- Lebeis, S. L., Rott, M., Dangl, J. L. & Schulze-Lefert, P. Culturing a plant microbiome community at the cross-Rhodes. *New Phytol.* **196**, 341–344 (2012).
- Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **108**, 6252–6257 (2011).
- Faure, D., Vereecke, D. & Leveau, J. J. Molecular communication in the rhizosphere. *Plant Soil* **321**, 279–303 (2009).
- Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S. & Vivanco, J. M. The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu. Rev. Plant Biol.* **57**, 233–266 (2006).
- Ramachandran, V. K., East, A. K., Karunakaran, R., Downie, J. A. & Poole, P. S. Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol.* **12**, R106 (2011).
- Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**, 996–998 (2013).
- Chi, F. *et al.* Ascending migration of endophytic rhizobia, from roots to leaves, inside rice plants and assessment of benefits to rice growth physiology. *Appl. Environ. Microbiol.* **71**, 7271–7278 (2005).

Supplementary Information is available in the online version of the paper.

**Acknowledgements** We thank D. Lundberg, S. Lebeis, S. Herrera-Paredes, S. Biswas and J. Dangl for sharing the calcined clay utilization protocol before publication; M. Kiselov of the ETH Zurich Flow Cytometry Core Facility for help with bacterial cell sorting as well as M. Baltisberger, D. Jolic and D. Weigel for their help in finding natural *Arabidopsis* populations; E. Kemen and M. Agler for sharing the Illumina Mi-Seq protocol for profiling of defined communities before publication and A. Sczyrba for his advice with the genome assembly. This work was supported by funds to P.S.-L. from the Max Planck Society, a European Research Council advanced grant (ROOTMICROBIOTA), the ‘Cluster of Excellence on Plant Sciences’ program funded by the Deutsche Forschungsgemeinschaft, the German Center for Infection Research (DZIF), by funds to J.A.V. from ETH Zurich (ETH Research Grant ETH-41 14-2), a grant from the Swiss National Research Foundation (310030B\_152835), and a European Research Council advanced grant (PhyMo).

**Author Contributions** J.A.V. and P.S.-L. initiated, coordinated and supervised the project. Y.B., M.R., N.D. and S.S. isolated root and soil bacteria strains. Y.B. collected root material and performed culture-independent community profiling. D.B.M., E.P. and M.R.-E. collected environmental leaf material, D.B.M. and E.P. isolated leaf strains and performed culture-independent community profiling. G.S. and R.G.-O. analysed culture-independent 16S rRNA amplicon sequencing data. Y.B., D.B.M. isolated DNA and prepared samples for genome sequencing. R.G.-O., P.C.M., B.H. and A.C.M. organized the genome sequencing data. R.G.-O. assembled and annotated draft genomes and performed comparative genome analyses. Y.B. and D.B.M. performed recolonization experiments; G.S. and R.G.-O. analysed the recolonization data. Y.B., D.B.M., R.G.-O., J.A.V. and P.S.-L. wrote the manuscript.

**Author Information** Sequencing reads (454 16S rRNA, MiSeq 16S rRNA and WGS HiSeq reads) have been deposited in the European Nucleotide Archive (ENA) under accession numbers PRJEB11545, PRJEB11583 and PRJEB11584, and genome assemblies and annotations corresponding to the leaf, root and soil culture collections have been deposited in the BioProject database under accession numbers PRJNA297956, PRJNA297942 and PRJNA298127. Isolates have been deposited at the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures (<https://www.dsmz.de/>). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence should be addressed to J.A.V. ([jvorholt@ethz.ch](mailto:jvorholt@ethz.ch)) or P.S.-L. ([schlef@mpipz.mpg.de](mailto:schlef@mpipz.mpg.de)).



## METHODS

**Sampling of *A. thaliana* plants and isolation of root-, leaf- and soil-derived bacteria.**

*A. thaliana* plants were either harvested from natural populations or grown in different natural soils and used for bacterial isolations by colony picking, limiting dilution or bacterial cell sorting as well as 16S rRNA gene-based community profiling. To obtain a library of representative root colonizing bacteria, *A. thaliana* plants were grown in different soils (50.958 N, 6.856 E, Cologne, Germany; 52.416 N, 12.968 E, Golm, Germany; 50.982 N, 6.827 E, Widdersdorf, Germany; 47.941 N, 04.012 W, Saint-Evarzec, France; 48.725 N, 3.989 W, Roscoff, France) and harvested before bolting. Briefly, *Arabidopsis* roots were washed twice in washing buffers (10 mM MgCl<sub>2</sub> for limiting dilution and PBS for colony picking<sup>6</sup>) on a shaking platform for 20 min at 180 rpm and then homogenized twice by Precellys24 tissue lyser (Bertin Technologies) using 3 mM metal beads at 5,600 rpm for 30 s. Homogenates were diluted and used for isolation approaches on several bacterial growth media (Supplementary Data 7). For isolations based on colony picking, diluted cell suspensions were plated on solidified media and incubated, before isolates of plates containing less than 20 colony-forming units (CFUs) were picked after a maximum of two weeks of incubation. For limiting dilution, homogenized roots from each root pool were sedimented for 15 min and the supernatant was empirically diluted, distributed and cultivated in 96-well microtitre plates<sup>20</sup>. In parallel to the isolation of root-derived bacteria, roots of plants grown in Cologne soil were harvested and used to assess bacterial diversity by culture-independent 16S rRNA gene sequencing. Additionally, soil-derived bacteria were extracted from unplanted Cologne soil by washing soil with PBS buffer, supplemented with 0.02% Silwet L-77 and subjected to bacterial isolation as well as 16S rRNA gene community profiling. For the isolation of representative phyllosphere strains, naturally grown *Arabidopsis* plants were collected at eight different sites in southern Germany and Switzerland (six main sampling sites used for bacterial isolations and community profiling: 47.4090306 N, 8.470169444 E, Hoengg, Switzerland; 47.474825 N, 8.305008333 E, Baden, Switzerland; 47.4816806 N, 8.217547222 E, Brugg, Switzerland; 48.5560194 N, 9.134944444 E, Farm, Tuebingen, Germany; 48.5989861 N, 9.201655556 E, Haeslach, Germany; 48.602682 N, 9.213247258 E, Haeslach, Germany; and two additional sites only used for bacterial isolation: 47.4074722 N, 8.50825 E, Zurich, Switzerland; 47.4227222 N, 8.548666667 E, Seebach, Switzerland) during spring and autumn of 2013 and used for bacterial isolations as well as 16S rRNA gene profiling. Leaf-colonizing bacteria of individual leaves were washed off by alternating steps of intense mixing and sonication. The suspension was subsequently filtered (CellTrics filters, 10 µm, Partec GmbH, Gölitz, Germany) in order to remove remaining plant or debris particles as well as cell aggregates and applied to cell sorting on a BD FACS Aria III (BD Biosciences) as well as to plating on different media (Supplementary Data 1 and 7). All isolates were subsequently stored in 30% or 40% glycerol at -80 °C.

**Culture-independent bacterial 16S rRNA gene profiling of *A. thaliana* leaf, root and corresponding soil samples.** Parts of *A. thaliana* leaves, roots and corresponding unplanted soil samples used for bacterial isolation were also processed for bacterial 16S rRNA gene community profiling using 454 pyrosequencing. Frozen root and corresponding soil samples were homogenized, DNA was extracted with Lysing Matrix E (MP Biomedicals) at 5,600 rpm for 30 s, and DNA was extracted from all samples using the FastDNA SPIN Kit for soil (MP Biomedicals) according to the manufacturer's instructions. Lyophilized leaf samples were transferred into 2 ml microcentrifuge tubes containing one metal bead and subsequently homogenized twice for 2 min at 25 Hz using a Retsch tissue lyser (Retsch, Haan, Germany). Homogenized leaf material was resuspended in lysis buffer of the MO BIO PowerSoil DNA isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA, USA), transferred into lysis tubes, provided by the supplier, and DNA extraction was performed following the manufacturer's protocol. DNA concentrations were measured by PicoGreen dsDNA Assay Kit (Life technologies), and subsequently diluted to 3.5 ng µl<sup>-1</sup>. Bacterial 16S rRNA genes were subsequently amplified<sup>6</sup> using primers targeting the variable regions V5-V7 (799F<sup>26</sup> and 1193R<sup>6</sup>, Supplementary Data 7). Each sample was amplified in triplicate by two independent PCR mixtures (a total of 6 replicates per sample plus respective no template controls). PCR products of triplicate were subsequently combined, purified and subjected to 454 sequencing. Obtained sequences were demultiplexed as well as quality and length filtered (average quality score ≥25, minimum length 319 bp with no ambiguous bases and no errors in the barcode sequences allowed)<sup>27</sup>. High-quality sequences were subsequently processed using the UPARSE<sup>24</sup> pipeline and OTUs were taxonomically classified using the Greengenes database<sup>28</sup> and the PyNAVE<sup>29</sup> method.

**High-throughput identification of leaf-, root- and soil-derived bacterial isolates by 454 pyrosequencing.** We adopted a two-step barcoded PCR protocol<sup>20</sup> in combination with 454 pyrosequencing to define V5-V8 sequences of bacterial 16S rRNA genes of all leaf, root- and soil-derived bacterial (Supplementary Fig. 1). DNA of isolates was extracted by lysis of 6 µl of bacterial cultures in 10 µl of buffer I containing 25 mM NaOH, 0.2 mM EDTA, pH 12 at 95 °C for 30 min, before the

pH value was lowered by addition of 10 µl of buffer II containing 40 mM Tris-HCl at pH 7.5. Position and taxonomy of isolates in 96-well microtitre plates were indexed by a two-step PCR protocol using the degenerate primers 799F and 1392R containing well- and plate-specific barcodes (Supplementary Data 7) to amplify the variable regions V5 to V8. During the first step of PCR amplification, DNA from 1.5 µl of lysed cells was amplified using 2 U DSF-Taq DNA polymerase, 1 × complete buffer (both Bioron GmbH), 0.2 mM dNTPs (Life technologies), 0.2 µM of 1 of 96 barcoded forward primer with a 18-bp linker sequence (for example, A1\_454\_799F1\_PCR1\_wells; Supplementary Data 7) and 0.2 µM reverse primer (454B\_1392R) in a 25 µl reaction. PCR amplification was performed under the following conditions: DNA was initially denatured at 95 °C for 2 min, followed by 40 cycles of 95 °C for 30 s, 50 °C for 30 s and 72 °C for 45 s, and a final elongation step at 72 °C for 10 min. PCR products of each 96-well microtitre plate were combined and subsequently purified in a two-step procedure using the Agencourt AMPure XP Kit (Beckman Coulter GmbH, Krefeld, Germany) first, then DNA fragments were excised from a 1% agarose gel using the QIAquick Gel Extraction Kit (Qiagen). DNA concentration was measured by Nanodrop and diluted to 1 ng µl<sup>-1</sup>.

During the second PCR step, 1 ng of pooled DNA (each pool represents one 96-well microtitre plate) was amplified by 1.25 U PrimeSTAR HS DNA Polymerase, 1 × PrimeSTAR Buffer (both TaKaRa Bio S.A.S, Saint-Germain-en-Laye, France), 0.2 mM dNTPs (Thermo Fisher Scientific Inc.), 0.2 µM of 1 of 96 barcoded forward primer targeting the 18-bp linker sequence (for example, P1\_454\_PCR2; Supplementary Data 7) and 0.2 µM reverse primer (454B\_1392R) in a 50 µl reaction. The PCR cycling conditions were as follows. First, denaturation at 98 °C for 30 s, followed by 25 cycles of 98 °C for 10 s, 58 °C for 15 s and 72 °C for 30 s, and a final elongation at 72 °C for 5 min. PCR products were purified using the Agencourt AMPure XP Kit (Beckman Coulter GmbH) and QIAquick Gel Extraction Kit (Qiagen) as described for the purification of first step PCR amplicons. DNA concentration was determined by PicoGreen dsDNA Assay Kit (Life technologies) and samples were pooled in equal amounts. The final PCR product libraries were sequenced on the Roche 454 Genome Sequencer GS FLX+. Each sequence contained a plate-barcode, a well-barcode and V5-V8 sequences.

The sequences were quality filtered, demultiplexed according to well and plate identifiers<sup>27</sup>. OTUs were clustered at 97% similarity by UPARSE algorithm<sup>24</sup>. A nucleotide-based blast (v. 2.2.29) was used to align representative sequences of isolated OTUs to culture-independent OTUs and only hits ≥97% sequence identity covering at least 99% of the length of the sequences were considered.

**Preparation of *A. thaliana* leaf (At-LSPHERE), root (At-RSPHERE) and soil bacterial culture collections.** Based on representative sequences of OTUs from this as well as previously published culture-independent community analysis, bacterial CFUs in the culture collections with ≥97% 16S rRNA gene identity to root-, leaf- and soil-derived OTUs were purified by three consecutive platings on the respective solidified media before an individual culture was used to inoculate liquid cultures. These liquid cultures were used for validation by Sanger sequencing with both 799F and 1392R primers as well as for the preparation of glycerol stocks for the culture collections and for the extraction of genomic DNA for whole-genome sequencing. A total of 21 leaf-derived strains, previously described as phyllosphere bacteria<sup>8,9</sup>, were added to the At-LSPHERE collection although these were undetectable in the present culture-independent leaf community profiling.

**Preparation of bacterial genomic DNA for whole-genome sequencing.** To obtain high molecular weight genomic DNA of bacterial isolates in our culture collections, we used a modified DNA precipitation protocol and the Agencourt AMPure XP Kit (Beckman Coulter GmbH). For each bacterial liquid culture, cells were collected by centrifugation at 3,220g for 15 min, the supernatant removed and cells were resuspended in 5 ml SET buffer containing 75 mM NaCl, 25 mM EDTA, 20 mM Tris/HCl at pH 7.5. A total of 20 µl lysozyme solution (50 mg ml<sup>-1</sup>, Sigma) was added before the mixture was incubated for 30 min at 37 °C. Subsequently, 100 µl 20 mg ml<sup>-1</sup> proteinase K (Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany) and 10% SDS (Sigma-Aldrich Chemie GmbH) were added, mixed, and incubated by shaking every 15 min at 55 °C for 1 h. If bacterial cells were insufficiently lysed, remaining cells were collected at 3,220g for 10 min and homogenized using the Precellys24 tissue lyser in combination with lysing matrix E tubes (MP Biomedicals) at 6,300 rpm for 30 s. After cell lysis, 2 ml 5 M NaCl and 5 ml chloroform were added and mixed by inversion for 30 min at room temperature. After centrifugation at 3,220g for 15 min, 6 ml supernatant were transferred into fresh falcon tubes and 3.6 ml isopropanol were added and gently mixed. After precipitation at 4 °C for 30 min, genomic DNA was collected at 3,220g for 5 min, washed once with 1 ml 70% (v/v) ethanol, dried for 15 min at room temperature and finally dissolved in 250 µl elution buffer (Qiagen). 2 µl 4 mg ml<sup>-1</sup> RNase A (Sigma-Aldrich Chemie GmbH) was added to bacterial genomic DNA solution and incubated over night at 4 °C.

The genomic DNA was subsequently purified using the Agencourt AMPure XP Kit (Beckman Coulter GmbH) and analysed by agarose gel (1% (w/v))

electrophoresis. Concentrations were estimated based on loaded Lambda DNA Marker (GeneRuler 1kb Plus, Thermo Scientific) and approximately 1 µg of genomic DNA was transferred into micro TUBE Snap-Cap AFA Fibre vials (Covaris Inc., Woburn, MA, USA). DNA was sheared into 350 bp fragments by two consecutive cycles of 30 s (duty cycle: 10%, intensity: 4, cycle/burst: 200) on a Covaris S2 machine (Covaris, Inc.). The Illumina sequencing libraries were prepared according to the manual of NEBNext Ultra UltraTM DNA Library Prep Kit for Illumina (New England Biolabs, USA). Quality and quantity was assessed at all steps by capillary electrophoresis (Agilent Bioanalyser and Agilent TapeStation). Finally libraries were quantified by fluorometry, immobilized and processed onto a flow cell with a cBot (Illumina Inc., USA) followed by sequencing-by-synthesis with TruSeq v3 chemistry on a HiSeq2500 (Illumina Inc., USA).

**Genome assembly and annotation.** Paired-end Illumina reads were subjected to quality and length trimming using Trimmomatic v. 0.33<sup>30</sup> and assembled using two independent methods (A5<sup>31</sup> and SOAPdenovo<sup>32</sup> v. 20.1). In each case, the assembly with the smaller number of scaffolds was selected. Detailed assembly statistics for each sequenced isolate can be found in Supplementary Data 3 and 4. Identification of putative protein-encoding genes and annotation of the genomes were performed using GLIMMER v. 3.02<sup>33</sup>. Functional annotation of genes was conducted using Prokka v. 1.11<sup>34</sup> and the SEED subsystems approach using the RAST server API<sup>35</sup>. Additionally, annotation of KEGG Orthologue (KO) groups was performed by first generating HMM models for each KO in the database<sup>36,37</sup> the HMMER toolkit (v. 3.1b2)<sup>38</sup>. Next, we employed the HMM models to search all predicted ORFs using the hmmsearch tool, with an *E* value threshold of  $10 \times 10^{-5}$ . Only hits covering at least 70% of the protein sequence were retained and for each gene and the match with the lowest *E* value was selected.

**Analyses of phylogenetic diversity within sequenced isolates.** Each proteome was searched for the presence of the 31 well-conserved, single-copy, bacterial AMPHORA genes<sup>39</sup>, designed for the purpose of high-resolution phylogeny reconstruction of genomes. Subsequently, a concatenated alignment of these marker genes was performed using Clustal Omega<sup>40</sup> v. 1.2.1. Based on this multiple sequence alignment, a species tree was inferred using FastTree<sup>41</sup> v. 2.1, a maximum likelihood tool for phylogeny inference. Whole-genome taxonomic classification of sequenced isolates was conducted using taxator-tk<sup>42</sup>, a homology-based tool for accurate classification of sequences. Analyses of phylogenetic diversity were performed independently for each cluster based on pairwise tree distances between all isolates (Supplementary Data 5).

**Analyses of functional diversity between sequenced isolates.** Analyses of functional diversity between sequenced isolates were conducted by generating, for each genome in the data set, a profile of presence/absence of each KO group (or phyletic pattern). Subsequently, a distance measure based on the Pearson correlation of each pair of phyletic patterns was calculated, which allowed us to embed each genome as a data point in a metric space. PCoA was performed on this space of functional distances using custom scripts written in R. Pairwise functional distances within each family-level cluster was performed by calculating the average distance between all pairs of genomes belonging to each cluster. Finally, we calculated RAs of each functional category based on the percentage of annotated KO terms assigned to each category. Enrichment tests were performed to identify differentially abundant categories between groups of genomes based on their origin (root versus leaf and root versus soil) using the non-parametric Mann–Whitney Test (MWT). *P* values were corrected for multiple testing using the Bonferroni method, with a significance threshold  $\alpha = 0.05$ .

**Recolonization experiments of leaf-, root- and soil-derived bacteria on *Arabidopsis*.** Calcined clay<sup>16</sup>, an inert soil substitute, was washed with water, sterilized twice by autoclaving and heat-incubated until being completely dehydrated. *A. thaliana* Col-0 seeds were surface-sterilized with ethanol and stratified overnight at 4°C. Leaf-, root- and soil-derived bacteria of the culture collections were cultivated in 96-deep-well plates and subsequently pooled (in equal or unequal ratios) in order to prepare synthetic bacterial communities (SynComs) for inoculations below the carrying capacity of leaves and roots<sup>43,44</sup>. To inoculate SynComs into the calcined clay matrix, OD<sub>600</sub> was adjusted to 0.5 and 1 ml ( $\sim 2.75 \times 10^8$  cells) was added to 70 ml 0.5 × MS media (pH 7; including vitamins, without sucrose), and mixed with 100 g calcined clay in Magenta boxes ( $\sim 2.75 \times 10^6$  cells per gr calcined clay), directly before sowing of surface-sterilized seeds. Plants were grown at 22°C, 11 h light, and 54% humidity. Alive cell counts (CFUs) of root-associated bacteria by serial dilutions of root homogenates after seven weeks of co-incubation were  $1.4 \times 10^8 \pm 8.4 \times 10^7$  cells per gram root tissue. For leaf spray-inoculation of *A. thaliana* plants, bacterial SynComs were prepared as described above and adjusted to OD<sub>600</sub> 0.2, before the solution was diluted tenfold and 170 µl ( $\sim 1.87 \times 10^6$  cells) were sprayed into each magenta box containing four three-week-old plants using a TLC chromatographic reagent sprayer (BS124.000, Biostep GmbH, Jahnsdorf, Germany). The average volume per spraying event was determined by spraying repeatedly into 50 ml tubes and weighing before and after. All plants and

corresponding unplanted clay samples were harvested under sterile conditions after a total incubation period of seven weeks. All plants and corresponding unplanted clay samples were harvested under sterile conditions after a total incubation period of seven weeks. During harvest, leaves and roots of individual plants were carefully separated using sterilized tweezers and scissors to avoid cross-contamination and processed separately thereafter. All leaves being obviously contaminated with clay particles or touching the ground were carefully removed and omitted from further processing. Remaining aerial parts of four plants collected from one magenta box were combined and transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at  $-80^\circ\text{C}$  until used for DNA extraction. Roots from one Magenta box were pooled, washed twice in 5 ml PBS at 180 rpm for 20 min, dried on sterilized Whatman glass microfibre filters (GE Healthcare Life Sciences), transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at  $-80^\circ\text{C}$  until further processing. The corresponding unplanted clay samples were washed in 100 ml PBS supplemented with 0.02% Silwet L-77 at 180 rpm for 10 min, before particles were allowed to settle down for 5 min. The supernatant was collected by centrifugation at 3,220g for 15 min. The pellet was subsequently resuspended in 1 ml water, transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at  $-80^\circ\text{C}$ .

To prepare DNA for bacterial 16S rRNA gene-based community analysis, all samples were homogenized twice by Precellys24 tissue lyser (Bertin Technologies), DNA was extracted and concentrations were measured by PicoGreen dsDNA Assay Kit (Life technologies), before bacterial 16S rRNA genes were amplified by degenerate PCR primers (799F and 1193R) targeting the variable regions V5–V7 (Supplementary Data 7). Each sample was amplified in triplicate (plus respective no template control) in 25 µl reaction volume containing 2 U DFS-Taq DNA polymerase, 1 × incomplete buffer (both Bioron GmbH, Ludwigshafen, Germany), 2 mM MgCl<sub>2</sub>, 0.3% BSA, 0.2 mM dNTPs (Life technologies GmbH, Darmstadt, Germany), 0.3 µM forward and reverse primer and 10 ng of template DNA. After an initial denaturation step at 94°C for 2 min, the targeted region was amplified by 25 cycles of 94°C for 30 s, 55°C for 30 s and 72°C for 60 s, followed by a final elongation step of 5 min at 72°C. The three independent PCR reactions were pooled and the remaining primers and nucleotides were removed by addition of 20 U exonuclease I and 5 U Antarctic phosphatase (both New England BioLabs GmbH, Frankfurt, Germany) and incubated for 30 min at 37°C in the corresponding 1 × Antarctic phosphatase buffer. Enzymes were heat-inactivated and the digested mixture was used as template for the 2nd step PCR using the Illumina compatible primers B5-F and 1 of 96 differentially barcoded reverse primers (B5-1 to B5-96, Supplementary Data 7). All samples were amplified in triplicate for 10 cycles using identical conditions of the first-step PCR. Technical replicates of each sample were combined, run on a 1.5% (w/v) agarose gel and the bacterial 16S rRNA gene amplicons were extracted using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions. DNA concentration was subsequently measured using the PicoGreen dsDNA Assay Kit (Life technologies) and 100 ng of each sample were combined. Final amplicon libraries were cleaned twice using the Agencourt AMPure XP Kit (Beckman Coulter GmbH) and subjected to sequencing on the Illumina MiSeq platform using an MiSeq Reagent kit v3 following the 2 × 350 bp paired-end sequencing protocol (Illumina Inc. USA).

Forward and reverse reads were joined, demultiplexed and subjected to quality controls using scripts from the QIIME toolkit<sup>27</sup>, v. 180 (Phred  $\geq 20$ ). The resulting high quality sequences were further clustered at 97% sequence identity together with Sanger sequences of leaf, root and soil isolates using the UPARSE<sup>24</sup> pipeline as described above. Taxonomic assignments of representative sequences were performed as explained in the previous sections. OTUs only corresponding to one or more Sanger 16S rRNA gene sequence(s) of purified strains in the *At*-RSPHERE, *At*-LSPHERE or soil collection were selected and designated 'indicator OTUs'. The heat maps were generated using the ggplot2 R package.

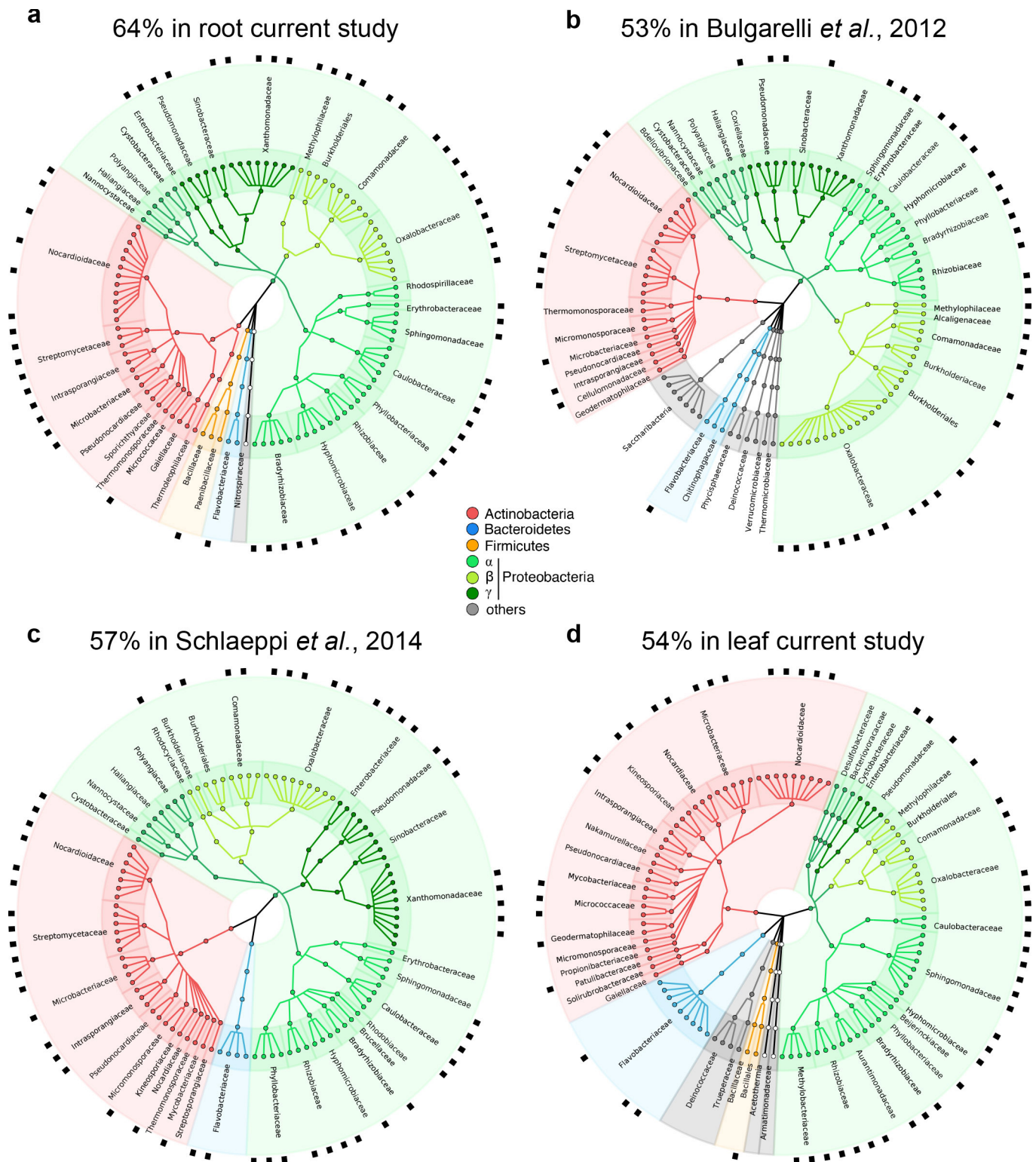
**Accession numbers.** Sequencing reads (454 16S rRNA, MiSeq 16S rRNA and WGS HiSeq reads) have been deposited in the European Nucleotide Archive (ENA) under accession numbers PRJEB11545, PRJEB11583 and PRJEB11584. Genome assemblies and annotations corresponding to the leaf, root and soil culture collections have been deposited in the National Center for Biotechnology Information (NCBI) BioProject database under accession numbers PRJNA297956, PRJNA297942 and PRJNA298127, respectively.

**Code availability.** All scripts for computational analysis and corresponding raw data are available at [http://www.mpi-pz.mpg.de/R\\_scripts](http://www.mpi-pz.mpg.de/R_scripts). The sequenced bacterial genomes as well as any future updates are available at <http://www.at-sphere.com>.

26. Chelius, M. K. & Triplett, E. W. The diversity of Archaea and Bacteria in association with the roots of *Zea mays* L. *Microb. Ecol.* **41**, 252–263 (2001).
27. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).

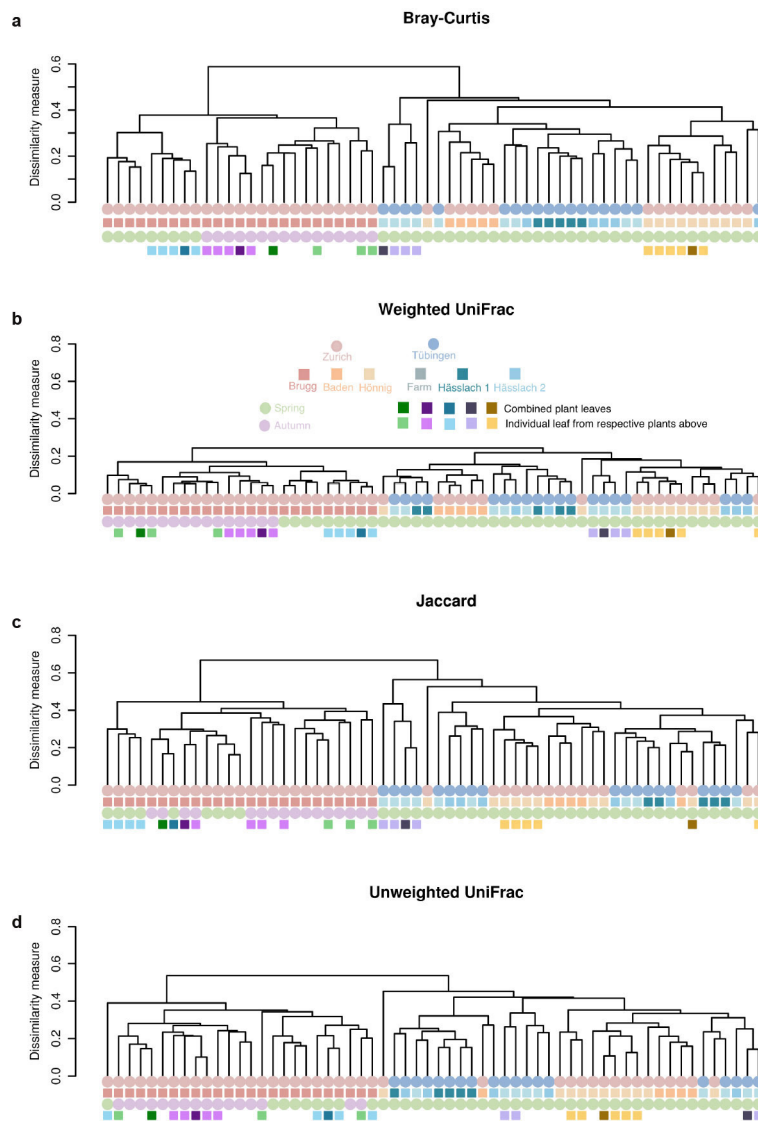
28. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
29. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267 (2010).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
31. Tritt, A., Eisen, J. A., Facciotti, M. T. & Darling, A. E. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One* **7**, e42304 (2012).
32. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
33. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
34. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
35. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
36. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
37. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
38. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
39. Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9**, R151 (2008).
40. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539–539 (2011).
41. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
42. Dröge, J., Gregor, I. & McHardy, A. C. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics* **31**, 817–824 (2015).
43. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
44. Bodenhausen, N., Bortfeld-Miller, M., Ackermann, M. & Vorholt, J. A. A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. *PLoS Genet.* **10**, e1004283 (2014).





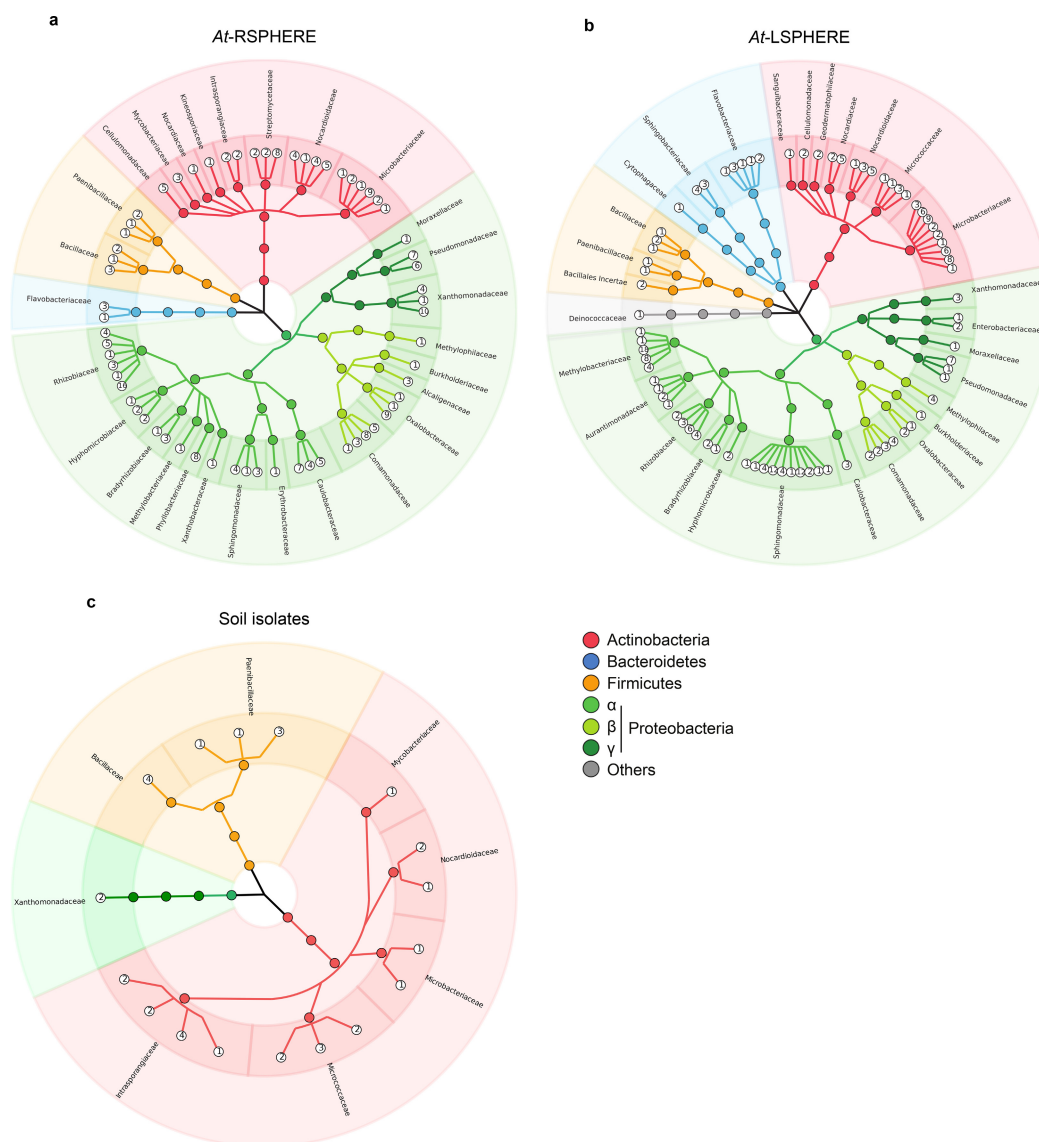
**Extended Data Figure 1 | Culture-dependent coverage of *A. thaliana* root- and leaf-associated OTUs identified in several cultivation-independent studies. a–d.** The inner circle depicts taxonomic assignments of top 100 root-associated OTUs (filled dots) for the indicated phyla and families that were identified in the current (a), ref. 6 (b) and

ref. 12 (c) studies with Cologne-soil-grown plants, and current leaf (d) study at locations around Tübingen and Zurich. Black squares of the outer ring highlight OTUs sharing  $\geq 97\%$  16S rRNA gene similarity to *Arabidopsis* root or leaf bacterial culture collection.



**Extended Data Figure 2 | 16S rRNA gene community profiling of phyllosphere samples from different locations. a–d,** The indicated Beta-diversity indices were calculated from leaf samples ( $n = 60$ ) collected

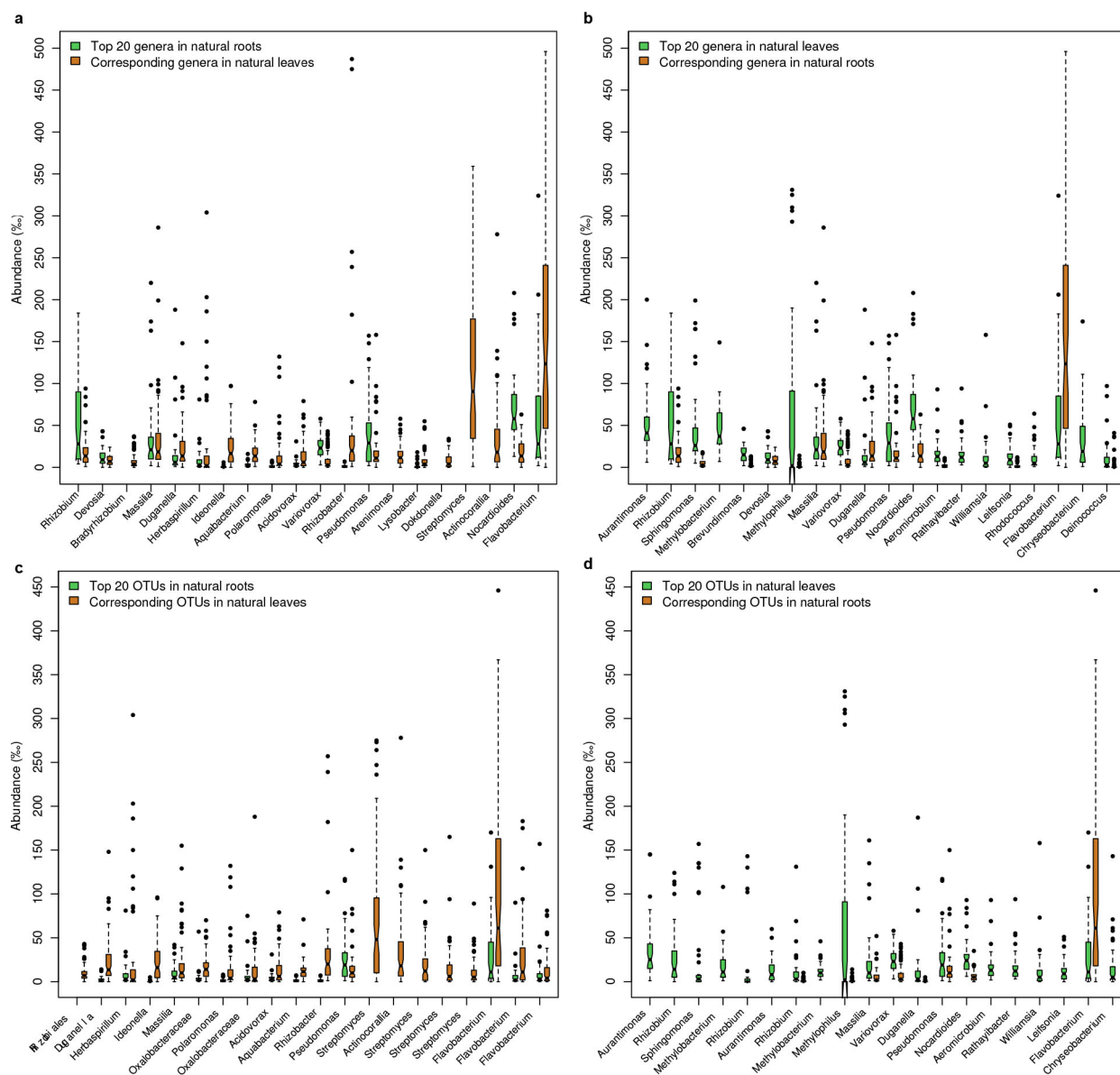
from natural *A. thaliana* populations growing in the areas around Tübingen and Zurich. The indicated colour code refers to sampling locations, sampling sites, sampling season, and combined or individual leaves of respective plants.



**Extended Data Figure 3 | *At*-RSPHERE, *At*-LSPHERE and soil bacterial culture collections.** **a**, *At*-RSPHERE ( $n = 206$  isolates), a culture collection of the *A. thaliana* root microbiota. **b**, *At*-LSPHERE ( $n = 224$  isolates), a culture collection of the *A. thaliana* leaf microbiota. **c**, Bacteria isolated

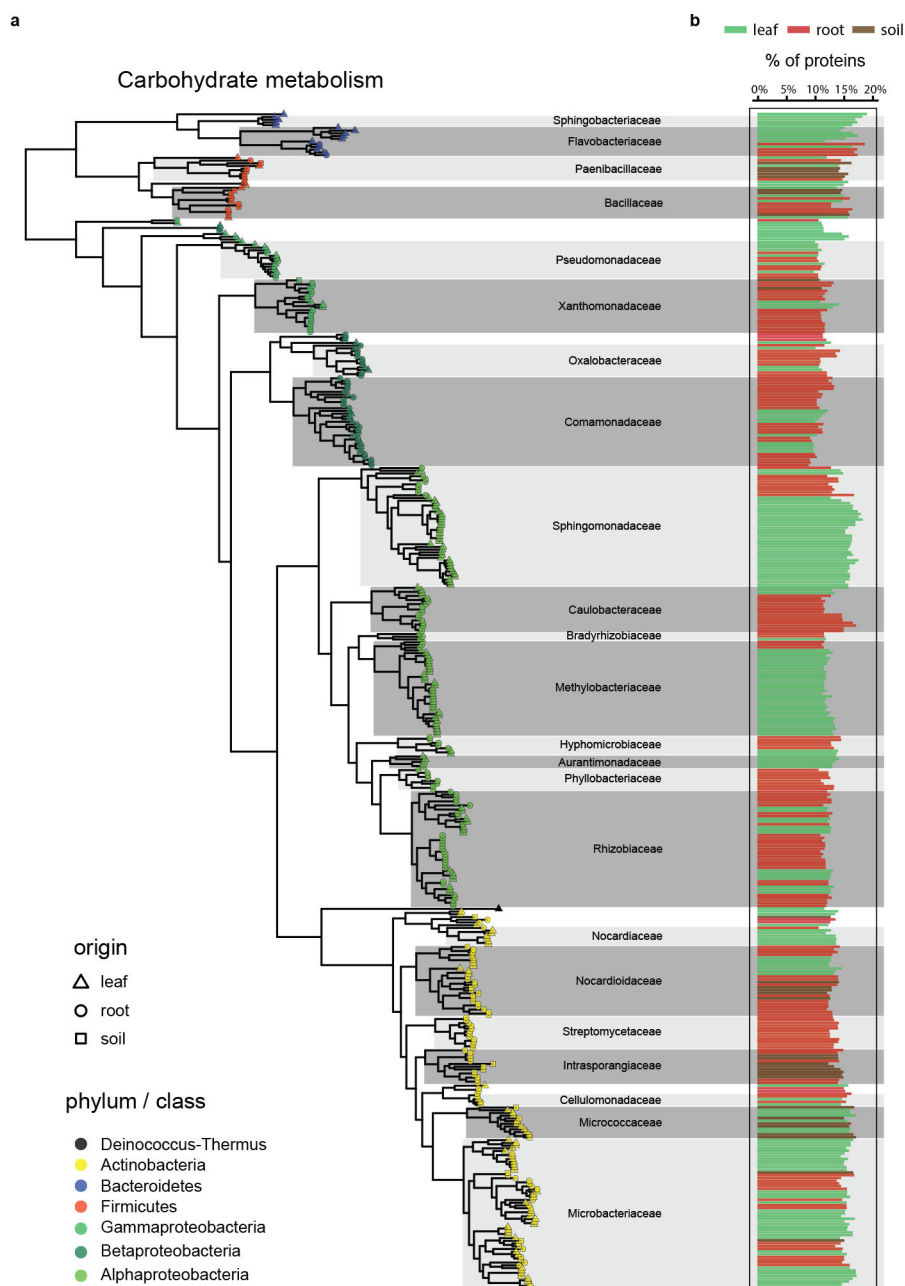
from Cologne soil ( $n = 33$  isolates). Numbers inside white circles indicate the number of bacterial isolates sharing  $\geq 97\%$  sequence identity, but isolated from independent roots, leaves and soil batches.





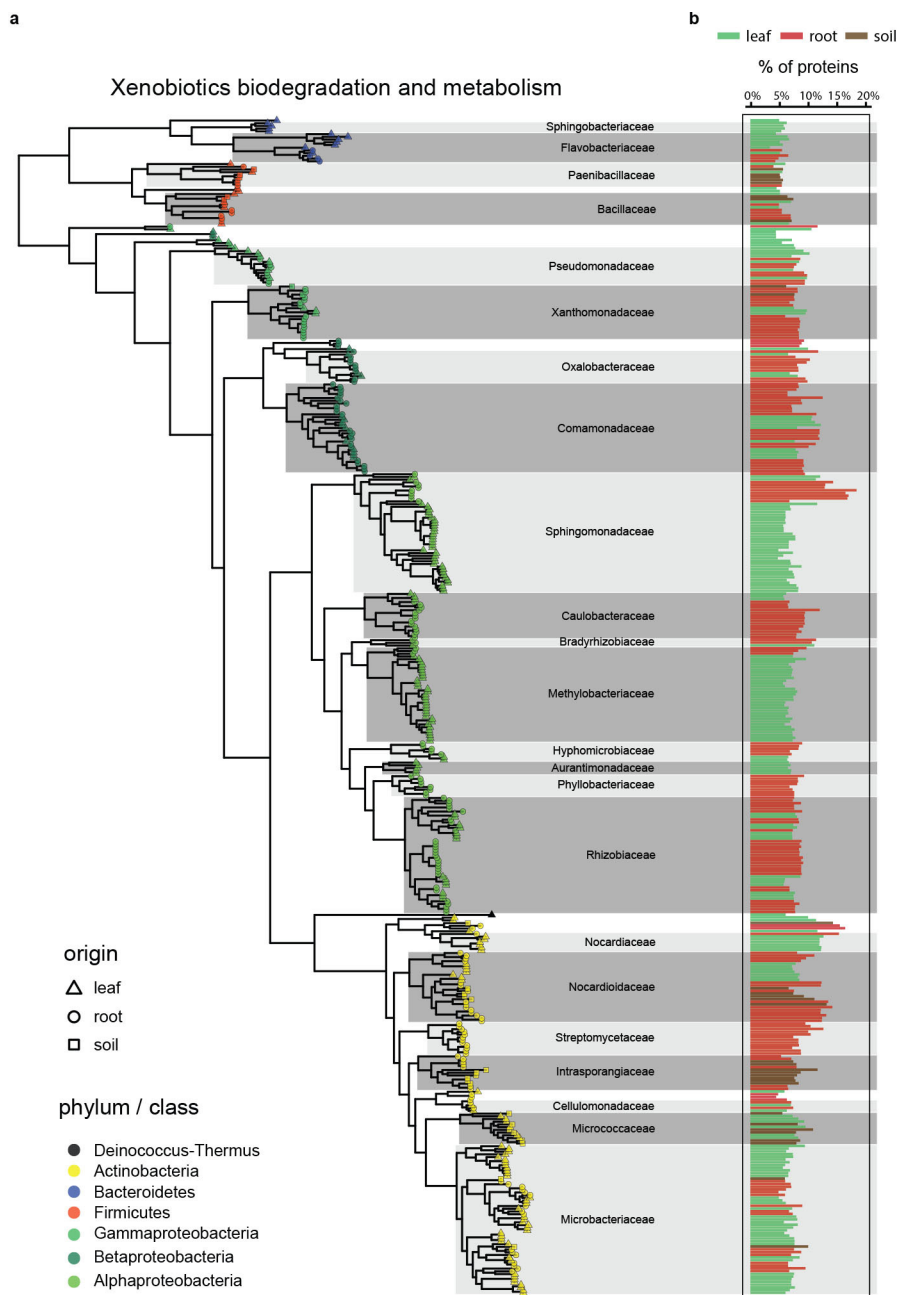
**Extended Data Figure 4 | Taxonomy overlap between *A. thaliana* root- and leaf-associated bacterial community from plants grown in natural soils.** a, b, Rank abundance plots of top 20 genera (a) and OTUs (b) in root bacterial communities ( $n=8$ ) from Cologne with corresponding genera detected in leaf bacterial communities ( $n=60$ ) from Zurich and

Tübingen. c, d, Rank abundance plots of top 20 genera (c) and OTUs (d) in leaf bacterial communities from Zurich and Tübingen with corresponding genera detected in root bacterial communities from Cologne. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.



**Extended Data Figure 5 | Phylogenetic distribution of 'carbohydrate metabolism' genes across sequenced isolates. a,** Phylogeny of sequenced leaf ( $n = 206$ ), root ( $n = 194$ ) and soil ( $n = 32$ ) isolates based on the concatenated alignment of the 31 conserved AMPHORA phylogenetic marker genes. The origin of each genome (leaf, root or soil) is shown by different shapes and their taxonomic affiliation (phylum level) is depicted

using various colours. Shaded areas correspond to the different clusters of genomes and are annotated with their consensus taxonomy (family level). **b,** Relative abundance of protein coding genes classified as belonging to the KEGG general term 'carbohydrate metabolism', measured as percentage of annotated proteins per genome.

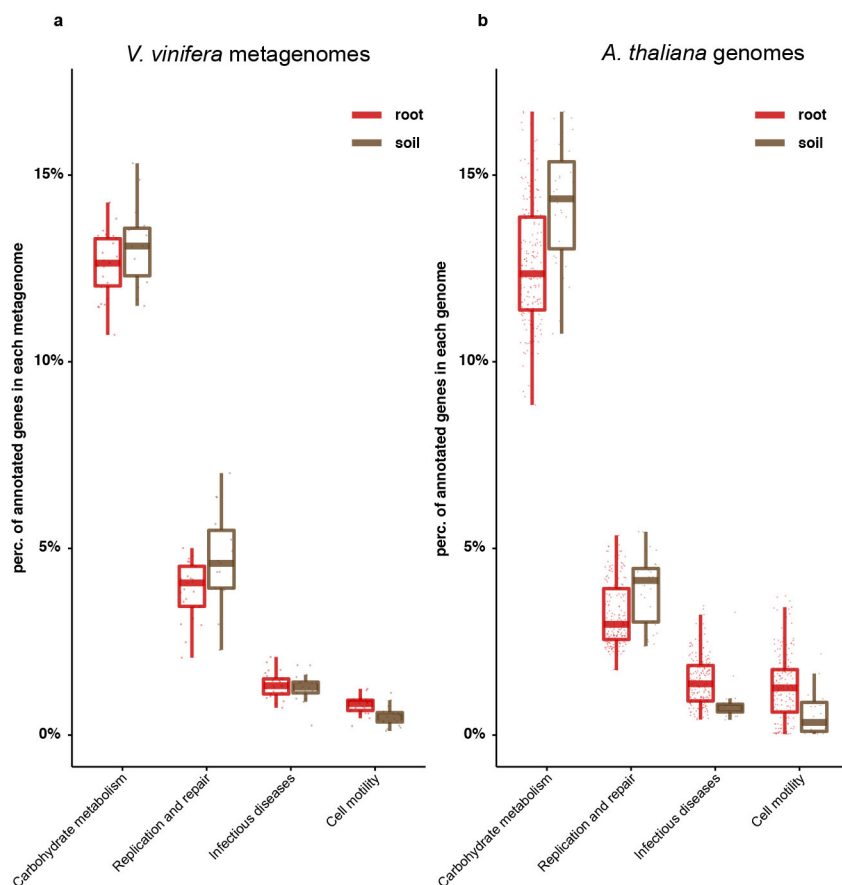


**Extended Data Figure 6 | Phylogenetic distribution of 'xenobiotic biodegradation and metabolism' genes across sequenced isolates.**

**a**, Phylogeny of sequenced leaf ( $n = 206$ ), root ( $n = 194$ ) and soil ( $n = 32$ ) isolates based on the concatenated alignment of the 31 conserved AMPHORA phylogenetic marker genes. The origin of each genome (leaf, root or soil) is shown by different shapes and their taxonomic

affiliation (phylum level; class level for Proteobacteria) is depicted using various colours. Shaded areas correspond to the different clusters of genomes and are annotated with their consensus taxonomy (family level). **b**, Relative abundance of protein coding genes classified as belonging to the KEGG general term 'xenobiotics biodegradation and metabolism', measured as percentage of annotated proteins per genome.

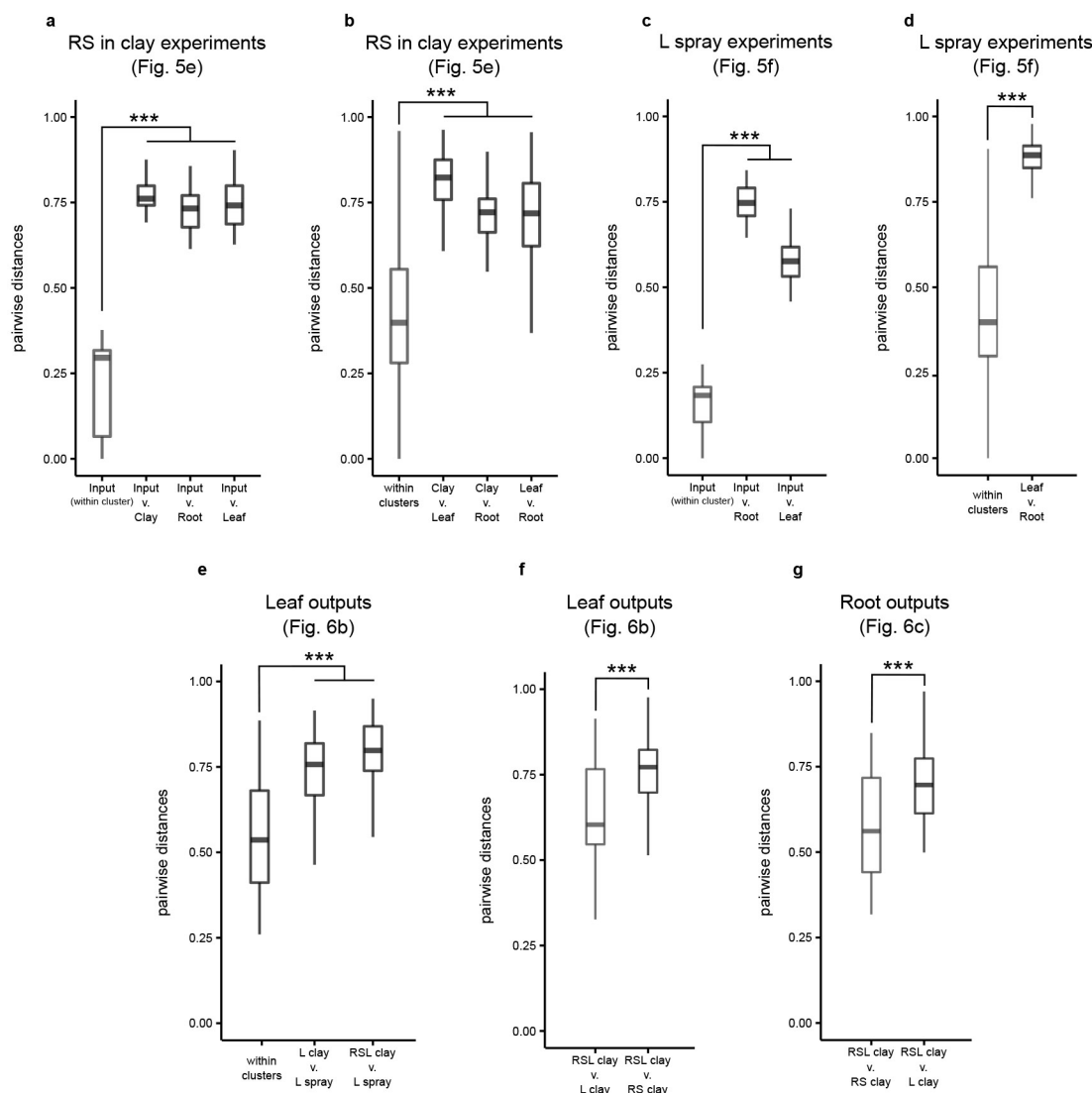




#### Extended Data Figure 7 | *V. vinifera* metagenome comparison.

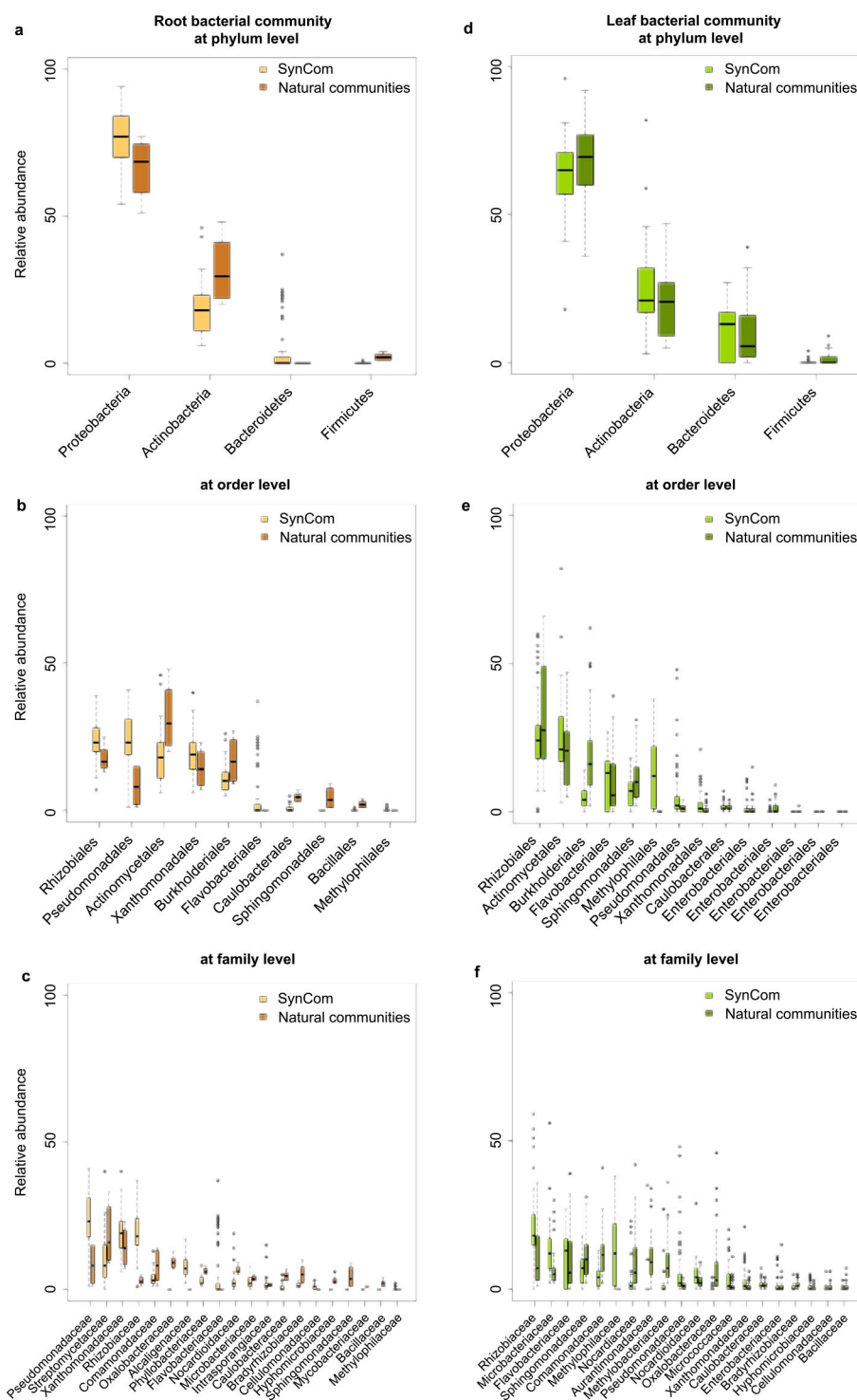
**a, b**, Functional enrichment analysis of *V. vinifera* root and soil shotgun metagenomes (**a**;  $n = 47$ ) compared to *A. thaliana* culture collection root and soil genomes (**b**;  $n = 432$ ). Functional category abundances

correspond to the percentage of annotated genes in each genome or metagenome sample. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.



**Extended Data Figure 8 | Cluster analysis of Bray–Curtis distances between groups of samples in the SynCom colonization of germ-free *A. thaliana* experiments.** **a**, Comparison of pairwise distances within input samples and between input and output samples of the RS in clay experiments. **b**, Comparison of pairwise distances between samples within the same cluster and between different clusters of the RS in clay experiments. **c**, Comparison of pairwise distances between input samples and between input and output samples of the L spray experiments. **d**, Comparison of pairwise distances within samples within the same cluster and between different clusters of the L spray experiments. **e**, Comparison of pairwise distances between samples within the same cluster and between different clusters of the leaf output across experiments.

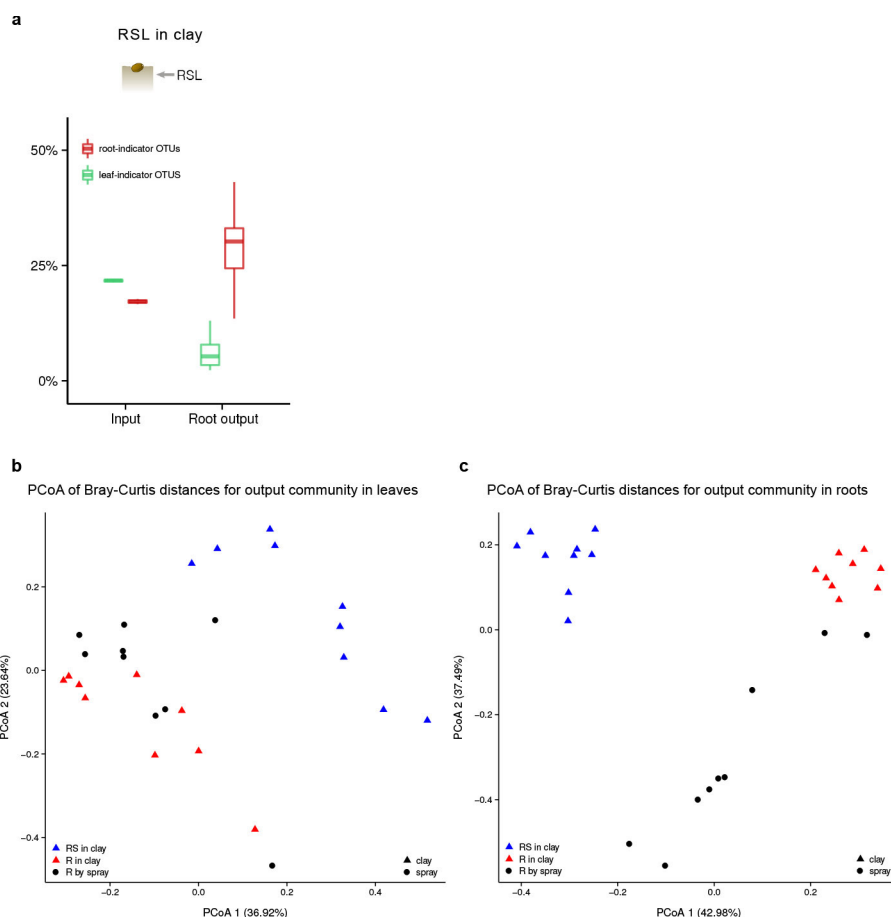
**f**, Comparison of pairwise distances between leaf output samples in the RSL in clay experiments and leaf output samples in the L in clay and RS in clay experiments. **g**, Comparison of pairwise distances between root output samples in the RSL in clay experiments and root output samples in the L in clay and RS in clay experiments. All comparisons marked with asterisks were subjected to a Student's *t*-test ( $P < 0.001$  in each case). L in clay was tested with 6 independently prepared SynComs ( $n = 6$ ); RSL in clay experiment was tested with 3 independently prepared SynComs, each used for 3 independent inoculations ( $n = 9$ ). All other experiments were tested with 6 independently prepared SynComs and each preparation was used for 3 independent inoculations ( $n = 18$ ). L, leaf-derived strains; RS, root- and soil-derived strains.



**Extended Data Figure 9 | Similarity of rank abundances of SynCom outputs with corresponding root- and leaf-associated OTUs of plants grown in natural environments.** a–c, Rank abundance plots of SynCom root outputs ( $n = 69$ ) with corresponding root-associated OTUs in natural communities ( $n = 8$ ) from plants grown in the present study in Cologne soil at the taxonomic ranks of phylum (a), order (b) and

family (c). d–f, Rank abundance plots of SynCom leaf outputs ( $n = 69$ ) with corresponding leaf-associated OTUs in natural communities ( $n = 60$ ) from plants grown in the present study around Tuebingen or Zurich at the taxonomic ranks of phylum (d), order (e) and family (f). Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.





**Extended Data Figure 10 | Fractional contribution of *At*-LSPHERE and *At*-RPSHERE-specific OTUs and SynCom competition supports host organ-specific community assemblies.** **a**, Fractional contribution of *At*-LSPHERE and *At*-RPSHERE specific OTUs in the input, leaf and the root output communities in the 'RSL in clay' experiment ( $n = 9$ ).

**b, c**, PCoA of Bray-Curtis distances of root (**b**;  $n = 21$ ) and leaf (**c**;  $n = 21$ ) outputs of the 'R in clay', 'RS in clay', and 'R spray' SynCom experiments.

R, root-derived isolates; S, soil-derived isolates; L, leaf-derived isolates. RSL in clay experiment was tested with 3 independently prepared SynComs, each used for 3 independent inoculations. All other experiments were tested with 3 independently prepared SynComs and each preparation was used for 3 independent inoculations. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.

# Phosphorylation and linear ubiquitin direct A20 inhibition of inflammation

Ingrid E. Wertz<sup>1,2</sup>, Kim Newton<sup>3</sup>, Dhaya Seshasayee<sup>4</sup>, Saritha Kusam<sup>2†</sup>, Cynthia Lam<sup>2</sup>, Juan Zhang<sup>4</sup>, Nataliya Popovych<sup>2</sup>, Elizabeth Helgason<sup>2</sup>, Allyn Schoeffler<sup>2</sup>, Surinder Jeet<sup>4</sup>, Nandhini Ramamoorthi<sup>4</sup>, Lorna Kategaya<sup>1,2</sup>, Robert J. Newman<sup>5</sup>, Keisuke Horikawa<sup>6</sup>, Debra Dugger<sup>3</sup>, Wendy Sandoval<sup>7</sup>, Susmith Mukund<sup>8</sup>, Anuradha Zindal<sup>2</sup>, Flavius Martin<sup>4</sup>, Clifford Quan<sup>2</sup>, Jeffrey Tom<sup>2</sup>, Wayne J. Fairbrother<sup>2</sup>, Michael Townsend<sup>4</sup>, Søren Warming<sup>5</sup>, Jason DeVoss<sup>4</sup>, Jinfeng Liu<sup>9</sup>, Erin Dueber<sup>2</sup>, Patrick Caplazi<sup>10</sup>, Wyne P. Lee<sup>4</sup>, Christopher C. Goodnow<sup>11</sup>, Mercedes Balazs<sup>4</sup>, Kebing Yu<sup>7</sup>, Ganesh Kolumam<sup>5</sup> & Vishva M. Dixit<sup>3</sup>

**Inactivation of the *TNFAIP3* gene, encoding the A20 protein, is associated with critical inflammatory diseases including multiple sclerosis, rheumatoid arthritis and Crohn's disease. However, the role of A20 in attenuating inflammatory signalling is unclear owing to paradoxical *in vitro* and *in vivo* findings. Here we utilize genetically engineered mice bearing mutations in the A20 ovarian tumour (OTU)-type deubiquitinase domain or in the zinc finger-4 (ZnF4) ubiquitin-binding motif to investigate these discrepancies. We find that phosphorylation of A20 promotes cleavage of Lys63-linked polyubiquitin chains by the OTU domain and enhances ZnF4-mediated substrate ubiquitination. Additionally, levels of linear ubiquitination dictate whether A20-deficient cells die in response to tumour necrosis factor. Mechanistically, linear ubiquitin chains preserve the architecture of the TNFR1 signalling complex by blocking A20-mediated disassembly of Lys63-linked polyubiquitin scaffolds. Collectively, our studies reveal molecular mechanisms whereby A20 deubiquitinase activity and ubiquitin binding, linear ubiquitination, and cellular kinases cooperate to regulate inflammation and cell death.**

Debilitating autoimmune syndromes and inflammatory diseases are associated with inactivation of the *TNFAIP3* gene, which encodes the A20 protein<sup>1</sup>. Despite the well-validated role of A20 in attenuating inflammation, fundamental mechanistic questions regarding A20 function remain unanswered. For example, A20 inactivation enhances pro-survival signalling and promotes expression of proteins that antagonize cell death<sup>2,3</sup>, yet A20 deficiency is also reported to sensitize cells to TNF-induced death<sup>2,4,5</sup>. Furthermore, A20 contains an OTU-type deubiquitinase domain that, in cells, cleaves scaffolding Lys63 (K63)-linked polyubiquitin to disassemble inflammatory receptor signalling complexes including tumour necrosis factor receptor-1 (TNFR1); however, *in vitro* A20 cleaves K48-, but not K63-linked polyubiquitin<sup>6,7</sup>. Finally, the A20 ZnF4 motif binds ubiquitin and facilitates substrate ubiquitination, but clear demonstration of A20 ubiquitin ligase function is not established<sup>8–10</sup> (Extended Data Fig. 1a).

## A20 edits ubiquitination of TNFR1 and associated proteins

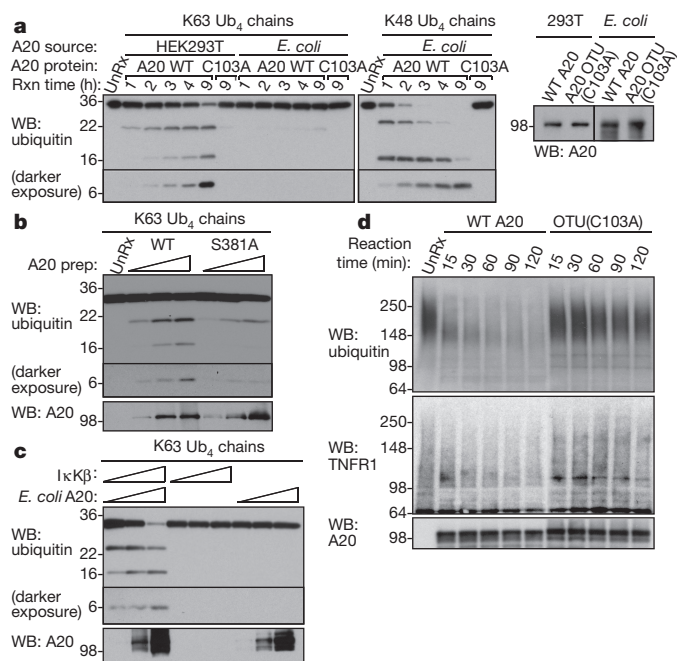
We performed proteomic, biochemical and *in vivo* analyses to investigate discrepancies regarding A20 function. To this end, we generated three strains of mice to compare the effect of inactivating A20 functional motifs *in vivo*: the OTU catalytic Cys103 was mutated to Ala to abolish A20 deubiquitinase activity in *Tnfaip3*<sup>OTU/OTU</sup> mice<sup>6,7,11</sup>, Cys609 and Cys612 were mutated to Ala to disrupt ZnF4 structure in *Tnfaip3*<sup>z4Cys/z4Cys</sup> mice<sup>8,9</sup>, and Tyr599 and Phe600 were mutated to Ala to compromise ZnF4 ubiquitin binding in *Tnfaip3*<sup>z4Ub/z4Ub</sup> mice<sup>9</sup> (Extended Data Fig. 2). *In vivo* studies confirmed heightened sensitivity of these mutant mice to TNF challenge (Extended Data Fig. 3a–d) and *Tnfaip3*<sup>OTU/OTU</sup> and *Tnfaip3*<sup>z4Cys/z4Cys</sup> mice had more severe myelin

oligodendrocyte glycoprotein-induced experimental autoimmune encephalomyelitis (MOG-EAE), a TNF-regulated disease model<sup>12</sup> (Extended Data Fig. 3e, f). These findings in *Tnfaip3*<sup>OTU/OTU</sup> mice and in *Tnfaip3*<sup>z4Cys/z4Cys</sup> mice agree with a previous study<sup>13</sup>. Importantly, the comparable phenotypes of *Tnfaip3*<sup>z4Cys/z4Cys</sup> and *Tnfaip3*<sup>z4Ub/z4Ub</sup> mice indicate that ZnF4 ubiquitin binding is critical for A20 function (Extended Data Fig. 3c, d). We next characterized signalling complexes in TNF-treated cells derived from *Tnfaip3*<sup>OTU/OTU</sup> and *Tnfaip3*<sup>z4Cys/z4Cys</sup> mice. Prolonged association of TNF receptor associated death domain (TRADD), transforming growth factor- $\beta$  activated kinase-1 (TAK1), and modified receptor interacting protein kinase-1 (RIPK1), but not I $\kappa$ B kinase- $\beta$  (I $\kappa$ B $\beta$ ), with activated TNFR1 was apparent in both types of A20 mutant cells relative to wild-type cells (Fig. 1a, b). Prolonged association of TAK1 and modified RIPK1 with activated TNFR1 corroborated enhanced downstream MKK3, MKK4, JNK and p38 MAPK activity, whereas transient I $\kappa$ B $\beta$  association was reflected in the modest enhancement of downstream NF $\kappa$ B signalling relative to wild-type cells (Fig. 1a, b, Extended Data Fig. 4d–g). A20 OTU(C103A), A20 ZnF4(C609A,C612A), and A20 ZnF4(Y599A,F600A) were all expressed and recruited to liganded TNFR1 at levels equivalent to or greater than wild-type A20, thereby excluding reduced association as an explanation for enhanced TNFR1 signalling (Extended Data Fig. 4a–c). We reasoned that the recruitment efficiency of TAK1 and I $\kappa$ B $\beta$  might reflect the global ubiquitination status of activated TNFR1 signalling components and thus serve as a functional readout of the effects of A20 OTU and ZnF4 domain inactivation. More specifically, because TAB2/3 in the TAK1 complex have the highest affinity for K63-linked polyubiquitin<sup>14–16</sup> the more pronounced differential in TAK1 recruitment between *Tnfaip3*<sup>OTU/OTU</sup> or *Tnfaip3*<sup>z4Cys/z4Cys</sup> cells

<sup>1</sup>Discovery Oncology, Genentech, South San Francisco, California 94080, USA. <sup>2</sup>Early Discovery Biochemistry, Genentech, South San Francisco, California 94080, USA. <sup>3</sup>Physiological Chemistry, Genentech, South San Francisco, California 94080, USA. <sup>4</sup>Immunology, Genentech, South San Francisco, California 94080, USA. <sup>5</sup>Molecular Biology, Genentech, South San Francisco, California 94080, USA. <sup>6</sup>Department of Cancer Biology and Therapeutics, The John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory 2601, Australia. <sup>7</sup>Protein Chemistry, Genentech, South San Francisco, California 94080, USA. <sup>8</sup>Structural Biology, Genentech, South San Francisco, California 94080, USA. <sup>9</sup>Bioinformatics, Genentech, South San Francisco, California 94080, USA. <sup>10</sup>Pathology, Genentech, South San Francisco, California 94080, USA. <sup>11</sup>Immunogenomics Laboratory, Immunology Division, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, New South Wales 2010, Sydney, Australia. <sup>†</sup>Present address: Gilead Sciences, Inc., Department of Biology, Foster City, California 94404, USA.







**Figure 2 | A20 phosphorylation promotes hydrolysis of K63-polyubiquitin chains.** **a**, Right, time course of K63- or K48-linked tetraubiquitin (Ub<sub>4</sub>) cleavage by wild-type or A20 OTU(C103A) proteins purified from *E. coli* or HEK293T cells. Left, unreacted A20 input samples. UnRx, unreacted. **b**, Cleavage of K63-linked-Ub<sub>4</sub> by wild-type or S381A A20 purified from HEK293T cells. **c**, Cleavage efficacy of K63-linked-Ub<sub>4</sub> by *E. coli*-derived, IκKβ-phosphorylated A20, IκKβ alone, or wild-type A20 alone. **d**, Time course of K63-linked-Ub<sub>4</sub> by *E. coli*-expressed, IκKβ-phosphorylated wild-type or A20 OTU(C103A). Ubiquitinated TNFR1 substrate was purified from Flag-TNF-treated *Tnfaip3*<sup>Otu/Otu</sup> MEFs. Gel source data are in Supplementary Figs 2, 3. Data represent two to five biological replicates.

K63-linked ubiquitin chains explains why A20 phosphorylation suppresses inflammatory signalling<sup>19</sup> and reveals a mechanism by which the activity of deubiquitinases for certain polyubiquitin chains can be modified by post-translational modifications.

### Linear polyubiquitin dictates the fate of A20-inactivated cells

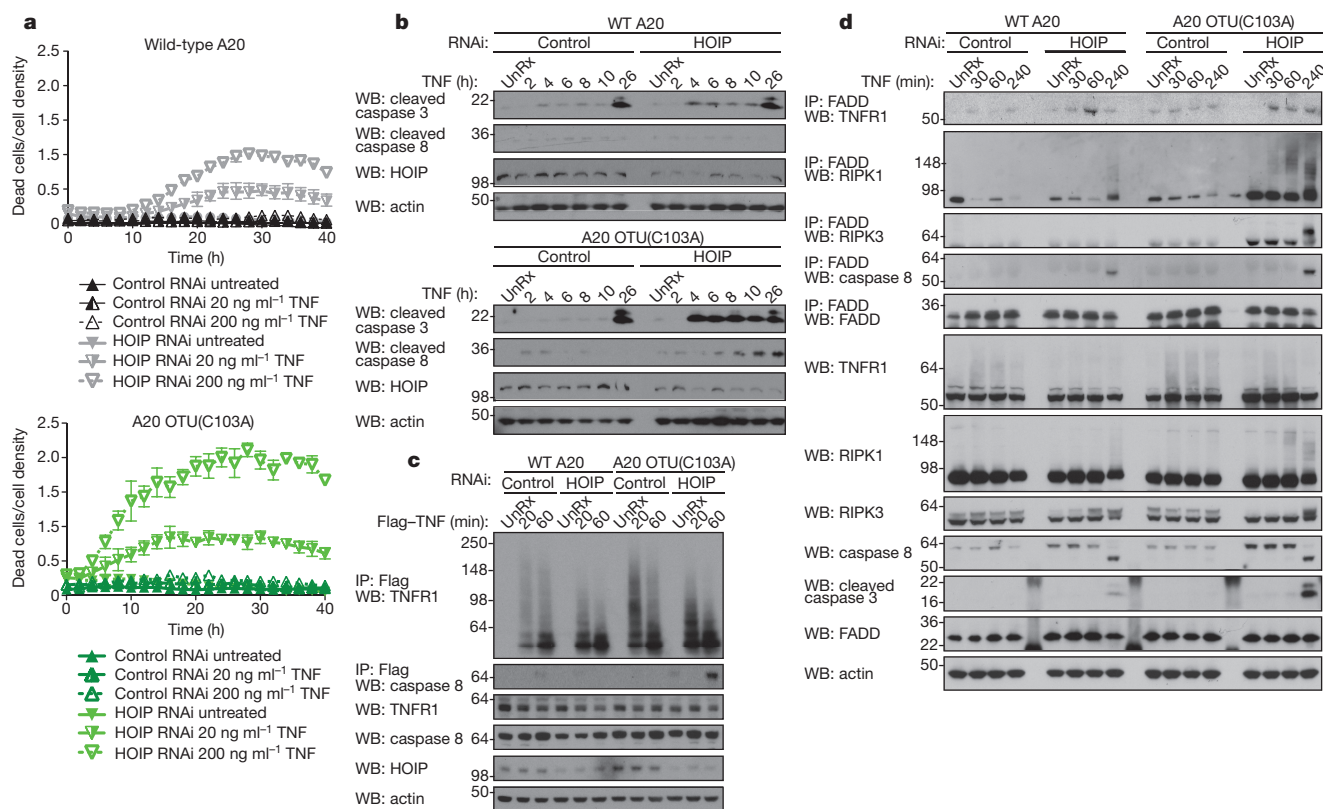
Our biochemical and cellular studies show that the A20 OTU domain depolymerizes K63-polyubiquitin chains (Fig. 2a–c, Extended Data Figs 5e, f and 6a, e) but not linear polyubiquitin (Extended Data Fig. 6c, g). Linear chains could, however, modulate A20 function. For example, A20 inactivation enhances pro-survival MAPK and NFκB signalling and promotes expression of proteins that antagonize cell death<sup>2,3</sup>, yet A20 deficiency paradoxically sensitizes cells to apoptotic and necroptotic TNF-induced death<sup>2,4,5</sup> and A20 expression protects against TNF-induced apoptosis<sup>20,21</sup>. However, the factors that determine whether A20 inactivation fosters cellular survival or demise are unknown. Because attenuated linear ubiquitination favours disassembly of the proximal TNFR1 signalling complex and promotes cell death<sup>22–28</sup>, we reasoned that linear chains might regulate whether compromised A20 function promotes cell survival or death. Knockdown of haem-oxidized IRP2 ubiquitin ligase-1 interacting protein (HOIP), a component of the linear ubiquitin chain assembly complex (LUBAC)<sup>29</sup>, reduced linear but not K63-linked ubiquitination of TNFR1 (Extended Data Fig. 7a) and, consistent with previous reports<sup>23–26</sup>, sensitized TNF-treated wild-type MEFs to death (Fig. 3a, Extended Data Fig. 7b). Importantly, *Tnfaip3*<sup>Otu/Otu</sup> MEFs were significantly more sensitive than wild-type MEFs to killing by TNF after HOIP knockdown (Fig. 3a, b, Extended Data Fig. 7b, active caspase quantitation not shown). Thus, linear ubiquitination dictates whether K63 hyperubiquitination resulting

from compromised A20 deubiquitinase activity promotes pro-inflammatory and cell survival signalling (Extended Data Fig. 4d, f) or cell death (Fig. 3a, Extended Data Fig. 7b). Supporting this notion, compromising TNFR1 K256 ubiquitination, that is primarily linear (Fig. 1c, Extended Data Fig. 5d), enhanced TNF-induced caspase activation (Extended Data Fig. 7c). Interestingly, compromised TNFR1 K256 ubiquitination did not affect JNK, p38, and NFκB signalling (Extended Data Fig. 7d), suggesting that TNFR1 ubiquitination regulates cell death whereas ubiquitination of TNFR1-associated proteins regulates downstream kinases. This idea corroborates a report that K63-linked ubiquitination of undefined TNFR1 residues regulates cell death but does not affect downstream kinase activity<sup>30</sup>. However, in this scenario attenuating TNFR1 ubiquitination blocked cell death—thus additional characterization of how TNFR1 ubiquitin modifications regulate signalling is warranted.

We profiled cellular signalling complexes to understand how A20 deubiquitinase activity and linear ubiquitination collaborate to regulate cell death. Caspase 8 is an apical protease in the TNF-induced cell death cascade<sup>31</sup> and its cleavage was detected earlier in *Tnfaip3*<sup>Otu/Otu</sup> MEFs compared to wild-type MEFs treated with HOIP RNA interference (RNAi; Fig. 3b), in keeping with more robust recruitment of caspase 8 to activated TNFR1 (Fig. 3c). Because engagement of TNFR1 is reported to promote recruitment of caspase 8 to a cytoplasmic “Complex II” rather than to TNFR1 at the plasma membrane (Complex I)<sup>32</sup>, we confirmed the specificity of our caspase 8 antibodies (Extended Data Fig. 7e) and analysed endogenous TNFR1 signalling complexes by mass spectrometry. Full-length caspase 8 was detected within the TNFR1 signalling complex along with receptor internalization components (Extended Data Figs 5a and 7f, g, and not shown). Thus caspase 8 may be recruited to TNFR1 complexes at the plasma membrane or to those that are endocytosed<sup>30</sup>; such recruitment is enhanced in cells with compromised A20 deubiquitinase activity and linear ubiquitination. The Fas-associated death domain (FADD) is an adaptor critical for assembly of both Complex I and Complex II<sup>31,32</sup>, and cleaved caspase 8 was most efficiently recruited to FADD in cells lacking both A20 deubiquitinase activity and linear ubiquitination, as was recruitment of RIPK3 and ubiquitinated RIPK1 (Fig. 3d). The presence of TNFR1 in anti-FADD immunoprecipitates substantiates interactions between membrane-bound and cytoplasmic cell death signalling components (Fig. 3d). Association of K63-ubiquitinated RIPK1 with the FADD-containing cell death complex was transient in *Tnfaip3*<sup>Otu/Otu</sup> MEFs but more robust in *Tnfaip3*<sup>Otu/Otu</sup> MEFs after HOIP knockdown, coinciding with enhanced caspase activation (Extended Data Fig. 7h). Thus enhanced K63-linked polyubiquitination of TNFR1 complex proteins resulting from A20 deficiency (Extended Data Figs 5e, f and 7h) facilitates assembly of either pro-survival or cell-death-inducing signalling complexes. Which complex prevails is dictated by linear ubiquitination: sufficient linear ubiquitination preserves the architecture of the pro-survival signalling complex for enhanced TAK1 and also IκKβ recruitment (Fig. 1a) and protracted downstream signalling (Extended Data Fig. 4d, f) whereas compromised linear ubiquitination favours association of TNFR1 and K63-hyperubiquitinated RIPK1 with cell death machinery and activation of cell death (Fig. 3, Extended Data Fig. 7b, c, f–h and diagram in Extended Data Fig. 1b, c).

### Linear ubiquitination prohibits A20 depolymerization of ubiquitin scaffolds

Next, we investigated how linear ubiquitination preserves the architecture of the TNFR1 pro-survival signalling complex<sup>22–28</sup>. A20 cannot depolymerize linear ubiquitin chains (Extended Data Figs 5f and 6c, g), thus linear ubiquitination could prohibit A20-mediated disassembly of the complex. Deubiquitinase profiling<sup>33</sup> confirmed that endogenous TNFR1 and RIPK1 are modified by K63-linked chains, that are depolymerized by A20, and linear chains, that are depolymerized by OTULIN<sup>34,35</sup> (Fig. 4a). The most complete deubiquitination was achieved by OTULIN and A20 (Fig. 4a). To evaluate whether linear



**Figure 3 | A20 regulates TNF-induced cell death in collaboration with linear ubiquitination.** **a**, Wild-type or *Tnfaip3*<sup>OTU/OTU</sup> MEFs transfected with control or HOIP RNAi oligonucleotides were treated with TNF. Ethidium homodimer-1-labelled dead cells were normalized to cell density. Mean values ( $n = 3$ )  $\pm$  s.e.m. **b**, wild-type or *Tnfaip3*<sup>OTU/OTU</sup> MEFs transfected with control or HOIP RNAi oligonucleotides were treated with TNF and cell lysates analysed by immunoblotting. **c**, Wild-type or *Tnfaip3*<sup>OTU/OTU</sup> control or HOIP RNAi-treated MEFs were

treated with Flag-TNF, engaged receptor complexes and cell lysates were analysed by immunoblotting. UnRx, untreated. **d**, Lysates from TNF-treated WT or *Tnfaip3*<sup>OTU/OTU</sup> control or HOIP RNAi-treated MEFs were immunoprecipitated using anti-FADD antibody; immunoprecipitates and cell lysates were immunoblotted as indicated. HOIP knockdown was validated by PCR with reverse transcription (RT-PCR; not shown). Gel source data are in Supplementary Figs 3, 4. Data represent two to four biological replicates.

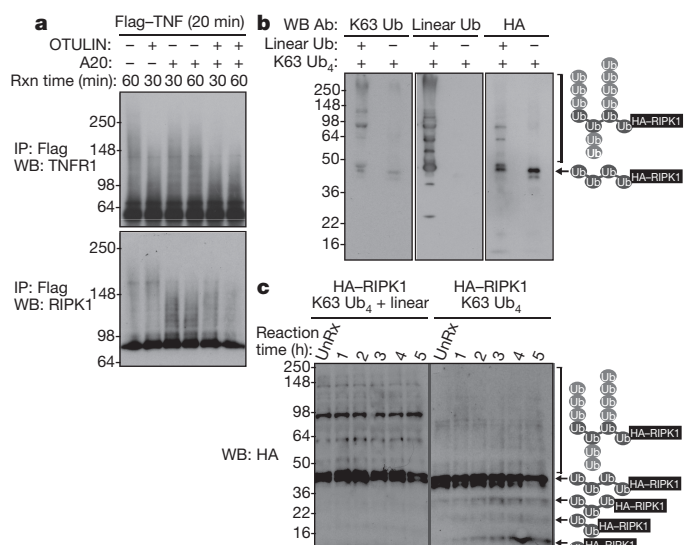
chains prohibit A20 from deubiquitinating K63-ubiquitinated TNFR1 signalling components, we polymerized linear ubiquitin upon a K63-tetraubiquitin chain that was ligated to an haemagglutinin-conjugated (HA)-RIPK1 peptide (Fig. 4b, Extended Data Fig. 6b). Whereas A20 depolymerized ubiquitin from K63-ubiquitinated HA-RIPK1, negligible deubiquitination occurred when the substrate was modified with branched linear chains (Fig. 4c, Extended Data Fig. 6a). Thus linear ubiquitination preserves the architecture of signalling complexes by physically prohibiting A20-mediated dissolution of K63 polyubiquitin scaffolds. Importantly, *Tnfaip3*<sup>OTU/OTU</sup> mice were also sensitized to lipopolysaccharide (LPS) challenge as reflected by enhanced lethality and elevated serum cytokines, and protracted MAPK and NF $\kappa$ B activation with persistence of K63-polyubiquitinated TRAF6 in isolated cells (Extended Data Fig. 8a, b, d). These data differ from a previous report, perhaps owing to treatment levels of LPS used<sup>36</sup>. Given that branched K63/linear chains are assembled on activated interleukin-1-receptor signalling components<sup>37</sup>, the mechanism by which linear ubiquitination prohibits deubiquitination of K63-polyubiquitin scaffolds and preserves the architecture of signalling complexes may be more broadly applicable to other signalling complexes regulated by linear ubiquitination and A20-like deubiquitinases<sup>38,39</sup>.

### A20 ZnF4 attenuates TNF signalling

Having investigated mechanisms by which A20 OTU activity regulates TNFR1 signalling, we next focused on the ZnF4 motif. A20 ZnF4 selectively binds K63-linked polyubiquitin chains<sup>9</sup>; nevertheless, both A20 ZnF4 mutants and wild-type A20 were recruited similarly to TNFR1 in *Tnfaip3*<sup>24Cys/24Cys</sup> and in *Tnfaip3*<sup>24Ub/24Ub</sup> cells (Fig. 5a, Extended Data Fig. 4b, c). This prompted us to define a recruitment mechanism

independent of ZnF4. A20 ZnF4 Cys residues support the ZnF4 structure<sup>9</sup>, and X-ray crystallography revealed that ZnF4 simultaneously binds to three monoubiquitins at ZnF4 sites I, II, and III (ref. 9). A20 ZnF7 also binds ubiquitin<sup>40,41</sup>, but the relative ubiquitin-binding affinities of ZnF4 and ZnF7 are unknown. We prepared nine individual <sup>15</sup>N-labelled human A20 ZnF proteins to quantify and compare directly monoubiquitin binding using NMR spectroscopy. ZnF7 bound monoubiquitin with a similar  $K_D$  to the ZnF4 ubiquitin-binding sites, and mutations in key ubiquitin-binding residues<sup>9,40–42</sup> suppressed binding (Extended Data Fig. 9a, Supplementary Information f–h). ZnF1 displayed negligible binding, thus ubiquitin binding is not a universal property of A20 zinc fingers (Extended Data Fig. 9a, Supplementary Information f). Because ZnF4 and ZnF7 monoubiquitin binding did not explain recruitment of A20 ZnF4(C609A,C612A) or A20 ZnF4(Y599A,F600A) to TNFR1, we used biolayer interferometry to measure binding affinities of A20 ZnF motifs to tri-ubiquitin chains. A20 ZnF7 bound to linear triubiquitin approximately 400 times more effectively than ZnF4 bound K63-linked triubiquitin (Fig. 5b, Extended Data Fig. 9a). Thus, in the absence of a functional ZnF4, ZnF7 is likely to direct A20 recruitment to the TNFR1 signalling complex via linear polyubiquitin binding. These data corroborate *TNFAIP3* mutational analyses in haematologic malignancies<sup>43</sup> and a role for ZnF7 in attenuating TNF-induced apoptosis<sup>44</sup>. ZnF7 mutants are probably not recruited to TNFR1, thus mimicking an A20-inactivated, TNF-sensitive phenotype. Supporting this idea, knockdown of HOIP reduced linear but not K63-linked ubiquitination of TNFR1 (Extended Data Fig. 7a) and attenuated A20 recruitment to TNFR1 (Fig. 5c). These data are in contrast to another study, which reported inefficient recruitment of murine A20 ZnF4(C609A,C612A) to activated TNFR1<sup>13</sup>. However,

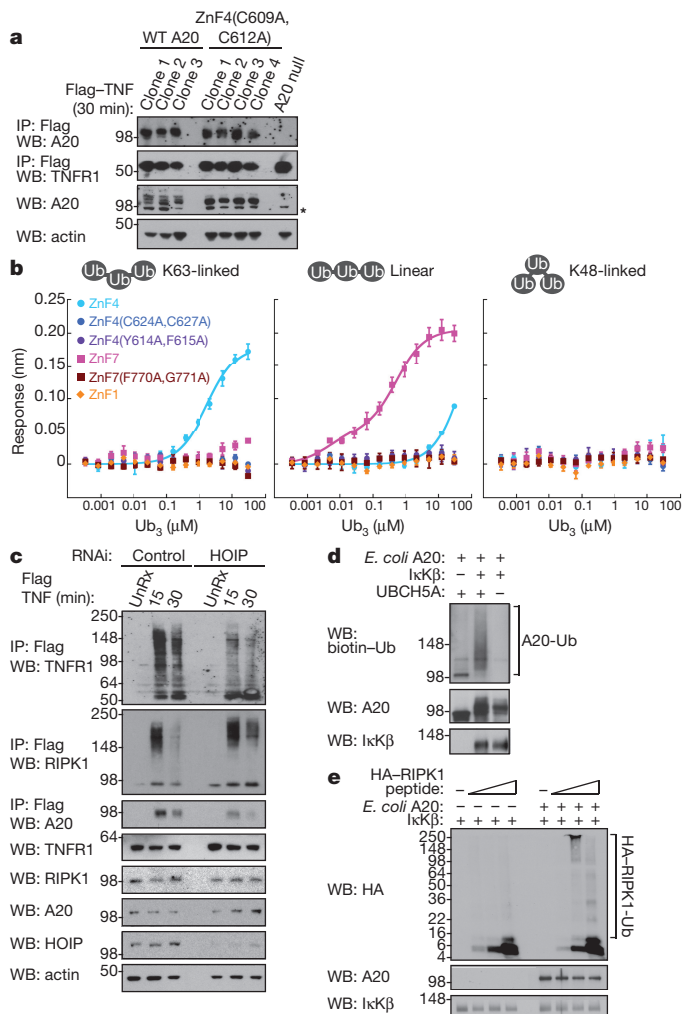




**Figure 4 | Linear ubiquitination prohibits A20 disassembly of the TNFR1 signalling complex.** **a**, Flag-TNF-purified elutions from *Tnfaip3*<sup>Otu/Otu</sup> MEFs were treated with recombinant deubiquitinases and immunoblotted as indicated. **b**, Immunoblot analysis of an HA-RIPK1 peptide modified with a K63-linked-Ub<sub>4</sub> chain or subsequently modified with linear ubiquitin chains. **c**, Cleavage of K63-linked-Ub<sub>4</sub> ligated to an HA-RIPK1 peptide with or without linear ubiquitination by *E. coli*-derived, IκKβ-phosphorylated A20. Gel source data are in Supplementary Figs 4, 5. Data represent two to three biological replicates.

the authors used anti-TNFR1 antibodies to isolate TNFR1, that do not effectively immunoprecipitate ubiquitinated TNFR1 (Extended Data Fig. 9b). Therefore, a significant fraction of the ubiquitinated and activated TNFR1 complex could have been inadvertently excluded from their analysis.

Effective A20 recruitment to activated TNFR1 is therefore insufficient for homeostatic A20 activity—the functional integrity of ZnF4 is also important. Although TNFR1-associated RIPK1 was increased in *Tnfaip3*<sup>z4Cys/z4Cys</sup> and *Tnfaip3*<sup>z4Ub/z4Ub</sup> cells (Fig. 1b, Extended Data Fig. 4c), K48-ubiquitinated RIPK1 associated with TNFR1 was decreased (Extended Data Fig. 9c). Thus, a functional A20 ZnF4 motif is required for K48-polyubiquitination of endogenous RIPK1. A20 ZnF4 also directed K48-linked polyubiquitination of TNFR1 *in vitro* (Extended Data Fig. 9d). Because levels of RIPK1 and TNFR1 are increased within the proximal TNFR1 signalling complex in *Tnfaip3*<sup>z4Cys/z4Cys</sup> and *Tnfaip3*<sup>z4Ub/z4Ub</sup> cells (Fig. 1b, Extended Data Fig. 4c) and K48-ubiquitin chains direct proteasomal degradation<sup>45</sup>, our data are consistent with a role for A20 ZnF4 in promoting RIPK1 and TNFR1 K48-linked ubiquitination and subsequent degradation to limit TNF signalling. Because IκKβ-mediated phosphorylation enhanced A20 K63 deubiquitinase activity (Fig. 2b–d, Extended Data Fig. 6a, e) we evaluated whether phosphorylation enhanced A20 ubiquitin ligase activity. IκKβ-phosphorylated A20 promoted more autoubiquitination than untreated A20 (Fig. 5d). IκKβ alone had no ligase activity (Extended Data Fig. 9e) and mass spectrometry of recombinant IκKβ did not reveal contaminating proteins of insect, viral or human origin that could mediate ligase or deubiquitinase function (not shown). Although biochemical and structural studies report A20 ZnF4 ubiquitin ligase activity<sup>8–10</sup>, it remains unclear whether *in vitro* A20 autoubiquitination is a phenomenon common to ubiquitin-binding motifs<sup>46</sup> or whether A20 ZnF4 has ubiquitin ligase function: the ability to transfer ubiquitin from a charged E2 enzyme to a substrate Lys residue and form polyubiquitin chains<sup>47</sup>. IκKβ-phosphorylated A20, but not IκKβ alone, transferred ubiquitin onto the Lys residue in the HA-RIPK1 peptide to generate polyubiquitinated substrate (Fig. 5e, Extended Data Fig. 6b). Thus the A20 ZnF4 motif has ubiquitin ligase activity that is enhanced by IκKβ-phosphorylation and is critical for attenuating



**Figure 5 | A20 ZnF4 ubiquitin-binding is required for attenuating TNF signalling.** **a**, Analysis of TNFR1-associated A20 from Flag-TNF-treated wild-type or *Tnfaip3*<sup>z4Cys/z4Cys</sup> MEFs. **b**, Equilibrium binding curves of human A20 ZnFs with Ub trimer (Ub<sub>3</sub>). Average response values ( $n = 3$ ) ± standard deviations. **c**, Wild-type MEFs transfected with control or HOIP RNAi oligonucleotides were treated with Flag-TNF; anti-Flag immunocomplexes or cell lysates were analysed by immunoblotting. **d**, Ubiquitination efficacy of *E. coli*-derived wild-type A20 or IκKβ-phosphorylated A20 with or without E2 UBCH5A. **e**, *E. coli*-derived, IκKβ-phosphorylated A20 promotes polyubiquitination of HA-RIPK1 peptide. Gel source data are in Supplementary Fig. 5. Data represent two to four biological replicates.

TNFR1 signalling. Intriguingly, A20 ZnF4 substrates and pathways affected by ZnF4 inactivation are selective: in contrast to *Tnfaip3*<sup>Otu/Otu</sup> mice, *Tnfaip3*<sup>z4Cys/z4Cys</sup> and *Tnfaip3*<sup>z4Ub/z4Ub</sup> mice were unaffected by LPS challenge (Extended Data Fig. 8c, d). Accordingly, cellular signalling and TRAF6 ubiquitination in *Tnfaip3*<sup>z4Cys/z4Cys</sup> cells remained unperturbed by LPS relative to wild-type cells (Extended Data Figs 1a, 8e).

## Discussion

Our studies have characterized physiological consequences of A20 OTU and ZnF4 domain inactivation. Because inactivating mutations in the OTU or ZnF4 domains are hypomorphic, simultaneous inactivation of OTU and ZnF4 domains, and/or the ZnF7 motif, may be required to fully incapacitate A20 and phenocopy *Tnfaip3*<sup>−/−</sup> mice. We find that A20 phosphorylation promotes K63-linked ubiquitin chain cleavage by the OTU domain and enhances ubiquitin transfer by the ZnF4 motif. Additional studies are required to understand how phosphorylation enhances A20 ubiquitin editing. While OTU and ZnF4 domains direct opposite enzymatic functions, TNF treatment yields



similar signalling profiles in A20 OTU- or ZnF4-domain mutant cells via two distinct mechanisms—more K63-ubiquitinated proteins accumulate in *Tnfaip3<sup>Otu/Otu</sup>* cells owing to insufficient deubiquitination, whereas insufficient degradation of K63-ubiquitinated proteins leads to their accumulation in *Tnfaip3<sup>z4Cys/z4Cys</sup>* and *Tnfaip3<sup>z4Ub/z4Ub</sup>* cells (Extended Data Fig. 1a).

We also show that TNFR1 ubiquitination and turnover are regulated by A20 OTU and ZnF4 domains. Stabilization of ligand-engaged TNFR1 could explain why *TNFAIP3* mutations are correlated with patient responses to anti-TNF therapies<sup>48</sup>: elevated levels of activated TNFR1 are a liability that can be exploited by TNF-neutralizing agents. Our finding that linear ubiquitination dictates whether A20 insufficiency enhances inflammatory signalling or cell death reveals an additional mechanism by which A20 regulates pathogenesis. Both outcomes can promote local or systemic inflammation *in vivo*<sup>22–28</sup>, underscoring why inactivating *TNFAIP3* mutations are associated with critical inflammatory and autoimmune syndromes<sup>1,43</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 5 November; accepted 11 October 2015.**

**Published online 9 December 2015.**

- Catrysse, L., Vereecke, L., Beyaert, R. & van Loo, G. A20 in inflammation and autoimmunity. *Trends Immunol.* **35**, 22–31 (2014).
- Lee, E. G. *et al.* Failure to regulate TNF-induced NF- $\kappa$ B and cell death responses in A20-deficient mice. *Science* **289**, 2350–2354 (2000).
- Kreuz, S., Siegmund, D., Scheurich, P. & Wajant, H. NF- $\kappa$ B inducers upregulate cFLIP, a cycloheximide-sensitive inhibitor of death receptor signaling. *Mol. Cell Biol.* **21**, 3964–3973 (2001).
- Vereecke, L. *et al.* Enterocyte-specific A20 deficiency sensitizes to tumor necrosis factor-induced toxicity and experimental colitis. *J. Exp. Med.* **207**, 1513–1523 (2010).
- Onizawa, M. *et al.* The ubiquitin-modifying enzyme A20 restricts ubiquitination of the kinase RIPK3 and protects cells from necroptosis. *Nature Immunol.* **16**, 618–627 (2015).
- Komander, D. & Barford, D. Structure of the A20 OTU domain and mechanistic insights into deubiquitination. *Biochem. J.* **409**, 77–85 (2008).
- Lin, S. C. *et al.* Molecular basis for the unique deubiquitinating activity of the NF- $\kappa$ B inhibitor A20. *J. Mol. Biol.* **376**, 526–540 (2008).
- Wertz, I. E. *et al.* De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF- $\kappa$ B signalling. *Nature* **430**, 694–699 (2004).
- Bosanac, I. *et al.* Ubiquitin binding to A20 ZnF4 is required for modulation of NF- $\kappa$ B signaling. *Mol. Cell* **40**, 548–557 (2010).
- Shembade, N., Ma, A. & Harhaj, E. W. Inhibition of NF- $\kappa$ B signaling by A20 through disruption of ubiquitin enzyme complexes. *Science* **327**, 1135–1139 (2010).
- Makarova, K. S., Aravind, L. & Koonin, E. V. A novel superfamily of predicted cysteine proteases from eukaryotes, viruses and *Chlamydia pneumoniae*. *Trends Biochem. Sci.* **25**, 50–52 (2000).
- Batoulis, H. *et al.* Blockade of tumour necrosis factor- $\alpha$  in experimental autoimmune encephalomyelitis reveals differential effects on the antigen-specific immune response and central nervous system histopathology. *Clin. Exp. Immunol.* **175**, 41–48 (2014).
- Lu, T. T. *et al.* Dimerization and ubiquitin mediated recruitment of A20, a complex deubiquitinating enzyme. *Immunity* **38**, 896–905 (2013).
- Kulathu, Y., Akutsu, M., Bremm, A., Hofmann, K. & Komander, D. Two-sided ubiquitin binding explains specificity of the TAB2 NZF domain. *Nature Struct. Mol. Biol.* **16**, 1328–1330 (2009).
- Sato, Y. *et al.* Structural basis for specific recognition of Lys 63-linked polyubiquitin chains by tandem UIMs of RAP80. *EMBO J.* **28**, 2461–2468 (2009).
- Husnjak, K. & Dikic, I. Ubiquitin-binding proteins: decoders of ubiquitin-mediated cellular functions. *Annu. Rev. Biochem.* **81**, 291–322 (2012).
- Rahighi, S. *et al.* Specific recognition of linear ubiquitin chains by NEMO is important for NF- $\kappa$ B activation. *Cell* **136**, 1098–1109 (2009).
- Hadian, K. *et al.* NF- $\kappa$ B essential modulator (NEMO) interaction with linear and lys-63 ubiquitin chains contributes to NF- $\kappa$ B activation. *J. Biol. Chem.* **286**, 26107–26117 (2011).
- Hutti, J. E. *et al.* I $\kappa$ B kinase beta phosphorylates the K63 deubiquitinase A20 to cause feedback inhibition of the NF- $\kappa$ B pathway. *Mol. Cell Biol.* **27**, 7451–7461 (2007).
- Daniel, S. *et al.* A20 protects endothelial cells from TNF-, Fas-, and NK-mediated cell death by inhibiting caspase 8 activation. *Blood* **104**, 2376–2384 (2004).
- Liuwantara, D. *et al.* Nuclear factor- $\kappa$ B regulates  $\beta$ -cell death: a critical role for A20 in beta-cell protection. *Diabetes* **55**, 2491–2501 (2006).
- Haas, T. L. *et al.* Recruitment of the linear ubiquitin chain assembly complex stabilizes the TNF-R1 signaling complex and is required for TNF-mediated gene induction. *Mol. Cell* **36**, 831–844 (2009).
- Gerlach, B. *et al.* Linear ubiquitination prevents inflammation and regulates immune signalling. *Nature* **471**, 591–596 (2011).
- Tokunaga, F. *et al.* SHARPIN is a component of the NF- $\kappa$ B-activating linear ubiquitin chain assembly complex. *Nature* **471**, 633–636 (2011).
- Ikeda, F. *et al.* SHARPIN forms a linear ubiquitin ligase complex regulating NF- $\kappa$ B activity and apoptosis. *Nature* **471**, 637–641 (2011).
- Peltzer, N. *et al.* HOIP deficiency causes embryonic lethality by aberrant TNFR1-mediated endothelial cell death. *Cell Reports* **9**, 153–165 (2014).
- Rickard, J. A. *et al.* TNFR1-dependent cell death drives inflammation in Sharpin-deficient mice. *eLife* **3**, e03464 (2014).
- Kumari, S. *et al.* Sharpin prevents skin inflammation by inhibiting TNFR1-induced keratinocyte apoptosis. *eLife* **3**, e03422 (2014).
- Walczak, H., Iwai, K. & Dikic, I. Generation and physiological roles of linear ubiquitin chains. *BMC Biol.* **10**, 23 (2012).
- Fritsch, J. *et al.* Cell fate decisions regulated by K63 ubiquitination of tumor necrosis factor receptor 1. *Mol. Cell Biol.* **34**, 3214–3228 (2014).
- Ashkenazi, A. & Dixit, V. M. Death receptors: signaling and modulation. *Science* **281**, 1305–1308 (1998).
- Micheau, O. & Tschopp, J. Induction of TNF receptor I-mediated apoptosis via two sequential signaling complexes. *Cell* **114**, 181–190 (2003).
- Mevisen, T. E. *et al.* OTU deubiquitinases reveal mechanisms of linkage specificity and enable ubiquitin chain restriction analysis. *Cell* **154**, 169–184 (2013).
- Rivkin, E. *et al.* The linear ubiquitin-specific deubiquitinase gumby regulates angiogenesis. *Nature* **498**, 318–324 (2013).
- Keusekotten, K. *et al.* OTULIN antagonizes LUBAC signaling by specifically hydrolyzing Met1-linked polyubiquitin. *Cell* **153**, 1312–1326 (2013).
- De, A., Dainichi, T., Rathinam, C. V. & Ghosh, S. The deubiquitinase activity of A20 is dispensable for NF- $\kappa$ B signaling. *EMBO Rep.* **15**, 775–783 (2014).
- Emmerich, C. H. *et al.* Activation of the canonical IKK complex by K63/M1-linked hybrid ubiquitin chains. *Proc. Natl Acad. Sci. USA* **110**, 15247–15252 (2013).
- Hitotsumatsu, O. *et al.* The ubiquitin-editing enzyme A20 restricts nucleotide-binding oligomerization domain containing 2-triggered signals. *Immunity* **28**, 381–390 (2008).
- Fiil, B. K. *et al.* OTULIN restricts Met1-linked ubiquitination to control innate immune signaling. *Mol. Cell* **50**, 818–830 (2013).
- Tokunaga, F. *et al.* Specific recognition of linear polyubiquitin by A20 zinc finger 7 is involved in NF- $\kappa$ B regulation. *EMBO J.* **31**, 3856–3870 (2012).
- Verhelst, K. *et al.* A20 inhibits LUBAC-mediated NF- $\kappa$ B activation by binding linear polyubiquitin chains via its zinc finger 7. *EMBO J.* **31**, 3845–3855 (2012).
- Skaug, B. *et al.* Direct, noncatalytic mechanism of IKK inhibition by A20. *Mol. Cell* **44**, 559–571 (2011).
- Ma, A. & Malynn, B. A. A20: linking a complex regulator of ubiquitylation to immunity and human disease. *Nature Rev. Immunol.* **12**, 774–785 (2012).
- Yamaguchi, N. & Yamaguchi, N. The seventh zinc finger motif of A20 is required for the suppression of TNF- $\alpha$ -induced apoptosis. *FEBS Lett.* **589**, 1369–1375 (2015).
- Komander, D. & Rape, M. The ubiquitin code. *Annu. Rev. Biochem.* **81**, 203–229 (2012).
- Hoeller, D. *et al.* E3-independent monoubiquitination of ubiquitin-binding proteins. *Mol. Cell* **26**, 891–898 (2007).
- Metzger, M. B., Pruneda, J. N., Klevit, R. E. & Weissman, A. M. RING-type E3 ligases: master manipulators of E2 ubiquitin-conjugating enzymes and ubiquitination. *Biochim. Biophys. Acta* **1843**, 47–60 (2014).
- Koczan, D. *et al.* Molecular discrimination of responders and nonresponders to anti-TNF $\alpha$  therapy in rheumatoid arthritis by etanercept. *Arthritis Res. Ther.* **10**, R50 (2008).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The authors thank J. Ernst, N. Kayagaki, K. O'Rourke, S. Hymowitz, I. Bosanac, E. Varfolomeev, T. Goncharov, M. diAlmagro, Z. Zhang, C. Klijn, D. Bustos, A. Maltzman, K. Wickliffe, J. Heideker, P. Liu, A. Ashkenazi, T.-K. Chang, B. Brasher, and C. Schwerdtfeger for reagents and discussions, A. Ma for A20 null MEFs, and the Genentech Protein Expression Group, Sequencing Facility, Luminex Core Group, Animal Facility and Genotyping Laboratory, J. Z. Solorio and M. Roose-Girma and the Mouse Models Group for dedication and support.

**Author Contributions** I.E.W., C.L., F.M., M.T., E.D., W.P.L., C.C.G., M.B. and V.M.D. coordinated studies. I.E.W., K.N., S.K., D.S., J.Z., N.P., E.H., A.S., S.J., N.R., L.K., K.H., D.D., W.S., A.Z., S.M., J.D.V., E.D., K.Y. and G.K. designed and performed experiments. I.E.W., K.N., D.S., J.Z., N.P., E.H., A.S., S.J., N.R., L.K., K.H., D.D., W.S., F.M., C.Q., W.J.F., M.T., S.W., J.D.V., J.L., E.D., P.C., W.P.L., C.C.G., M.B., K.Y., G.K. and V.M.D. interpreted data. I.E.W. wrote the manuscript. I.E.W., K.N., N.P., S.M., E.H., R.J.N., C.Q., J.T. and S.W. prepared reagents.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.E.W. ([ingrid@gene.com](mailto:ingrid@gene.com)) or V.M.D. ([dixit@gene.com](mailto:dixit@gene.com)).

## METHODS

**Generation of mice carrying the *Tnfaip3* C103A, C609A/C612A, or Y599A/F600A knock-in alleles.** The three constructs for targeting the *Tnfaip3* locus in embryonic stem (ES) cells were made using recombineering and/or standard molecular cloning techniques. The *Tnfaip3* C103A knock-in (KI) construct corresponds to the following genomic position (all coordinates from the NCBI37/mm9 assembly, complementary strand): chr10: 18725681–18733460. The C103A mutation (TGC to GCC) is in exon 3, and a loxP-Neo-loxP cassette was inserted in a position corresponding to chr10:18729002. The *Tnfaip3* C609A/C612A KI construct corresponds to chr10:18722957–18730583. The C609A/C612A double mutation (TGCCTCTATGT to GCTACTCTAGCT) is in exon 7, and a loxP-Neo-loxP cassette was inserted in a position corresponding to chr10: 18726067. The *Tnfaip3* Y599A/F600A KI construct corresponds to chr10: 18721314–18727086, the Y599A/F600A double mutation (TATTTT to GCTGCA) is located in exon 7, and an Frt-Neo-Frt selection cassette was inserted at a position corresponding to chr10: 18724150. The vectors were confirmed by DNA sequencing, linearized and used to target C57BL/6 C2 ES cells using standard methods (G418 positive and ganciclovir negative selection). Positive clones were identified using Southern blot and/or PCR, quantitative PCR, and sequencing. Correctly targeted ES cells were transfected with a Cre or Flpe plasmid, respectively, to remove the Neo selection marker and to create ES cells with the final *Tnfaip3* KI alleles. KI ES cells were injected into blastocysts using standard techniques, and germline transmission was obtained after crossing resulting chimaeras with C57BL/6N females.

**Genotyping wild-type and *Tnfaip3* KI mice.** Genomic DNA was extracted from tails using the Qiagen DNeasy Blood and Tissue Kit according to the manufacturer's instructions. Genomic DNA was amplified using the Invitrogen GeneAmp Fast PCR Master Mix following standard procedures.

Primer sequences and PCR products are as follows:

A20 OTU(C103A): Expected PCR product sizes: WT, 321 bp; KI.MUT, 383 bp.

Reverse primer: CCTCCAGTGCATTCCTGAGGAATCTC.

Forward primer: AAGCATGCACGATGAAGGAGC.

A20 ZnF4(C609A,C612A): Expected PCR product sizes: WT, 738 bp; KI.MUT, 873 bp.

Reverse primer: GCCTTGACAGGGATCTCCAT.

Forward primer: CACTCTCATGGTGTCTTCTGAGATG.

A20 ZnF4(Y599A,F600A): Expected PCR product sizes: WT, 420 bp; KI.MUT, 454 bp.

Primer1: TCTCACTCCACACTCTTG.

Primer2: TTCAGACCGAAGTTCCTAT.

Primer3: TGGGCTACATAATGGGTTTA.

**In vivo cytokine challenge studies and data analysis.** All staff participating in animal work abided by the laws and regulations as stated in the Animal Welfare Act and The Guide for the Care and Use of Laboratory Animals. This study did not unnecessarily duplicate previous experiments. Alternatives to the use of animals for this study were considered and none existed or were acceptable. All scientists and technicians were trained in the proper procedures for animal handling, animal techniques, the administration of anaesthesia and analgesia, and the methods of euthanasia used in these studies. All animal study protocols were reviewed and approved by the Genentech Institutional Animal Care and Use Committee (IACUC), and mice were generated and housed in a clean rodent facility on-site at Genentech Dixon and South San Francisco campuses in standard rodent micro-isolator cages. Sample size was generated based on historical data and knowledge of variability within the cytokine challenge models. Animals were selected based on body weight, age, and genotype and subsequently randomized based on body weight, with animals of extremes of body weight or age excluded from the study. The investigators were not blinded to allocation during experiments and outcome assessment.

For TNF challenge studies, 300 µg murine TNF per kg body weight (Genentech, Inc.) in PBS was injected intravenously. Female C57bl6 mice between 2–4 months of age were used for all studies with *Tnfaip3*<sup>OTU/OTU</sup> and *Tnfaip3*<sup>24Cys/24Cys</sup> animals. For studies with *Tnfaip3*<sup>24Cys/24Cys</sup> and *Tnfaip3*<sup>24Urb/24Urb</sup> animals, 9-week old male C57bl6 mice were used and dosing of TNF or PBS was done in a sequential manner by alternating wild-type and mutant mice and the samples are collected accordingly. Serum was collected at the indicated time points from three mice each of the indicated genotype for Luminex multiplex cytokine analysis. The rectal temperatures of 3–4 mice per genotype were also recorded at the indicated time points. Dosing of TNF or PBS was blinded and data were collected in a blinded manner.

LPS (Sigma L3012) in PBS was administered by intraperitoneal injection for challenge studies. Mice were challenged with 20 mg or 40 mg LPS per kg body weight for mortality studies, and both Log Rank (Mantel-Cox) and Wilcoxon tests were used to calculate *P* values.

For high-dose LPS challenge studies, serum was collected at the indicated time points from three mice each of the indicated genotype for Luminex multiplex cytokine analysis in response to 40 mg LPS per kg body weight. For low dose LPS challenge studies 4 or 5 mice per genotype were injected with PBS or 5 mg LPS per kg body weight in PBS. Serum was collected at 2 h or 6 h post-injection for Luminex multiplex cytokine analysis. A two-tailed Student's *t*-test was used to calculate *P* values.

**Antibodies and reagents.** Antibodies to the indicated proteins were purchased from the specified vendors, with catalogue or clone numbers indicated in brackets. Anti-FADD (Ab52935), anti-caspase 8 (Ab138485) (AbCam); anti-caspase 8 (ALX804448-C100) (Enzo); anti-RIPK1 (610459) (BD Biosciences); biotinylated anti-murine TNFR1 (BAF425), hamster anti-murine TNFR1 (55R170), anti-TAK1 (491840), anti-IκK3 (725818) (R&D systems); anti-A20 (5630), anti-IκBα (9242), anti-phospho-IκBα (5A5), anti-JNK (56G8), anti-phospho-JNK (81E11), anti-p38 (9212), anti-phospho-p38 (D3F9) anti-phospho-MKK3 (12280), anti-total MKK3 (8535), anti-phospho-MKK4 (4514), anti-total MKK4 (9152), anti-mouse-specific caspase 8 (4927) anti-cleaved caspase 8 (8592), anti-human caspase 8 (9746), anti-cleaved caspase 3 (9664), anti-human caspase 3 (9662), anti-PARP (9541) (Cell Signaling Technology); anti-β-tubulin (clone DM1B) (MP Biomedicals); anti-RIPK3 (NBP1-77299) (Novus); anti-TRAF2 (sc-7346), anti-TRAF6 (sc-7221), anti-actin-HRP (sc-1616), anti-ubiquitin and ubiquitin-HRP (clone P4D1), mouse anti-human TNFR1 (sc-8436), anti-TAK1 (sc-7162) (Santa Cruz Biotechnology); anti-HA-HRP (clone HA-7), anti-TRADD (SAB44503461), anti-RNF31 (HOIP) (SAB2102031), anti-Flag M2 affinity gel (A2220) (Sigma). Anti-K11 polyubiquitin, anti-K48 polyubiquitin, anti-K63 polyubiquitin, anti-linear polyubiquitin, anti-FADD (9274) and non-production grade Trastuzumab antibodies were produced at Genentech. Recombinant human TNF was produced at Genentech. Tri-ubiquitin and tetra-ubiquitin chains of indicated linkages were generated at Genentech. Recombinant human Flag-TNF was purchased from Enzo Life Sciences or was produced at Genentech. LPS was purchased from Sigma and recombinant murine GM-CSF and M-CSF were purchased from R&D systems.

**Cell culture.** MEFs were generated from E14 embryos and in some cases were immortalized by retroviral transduction with pWZL-hygro E1A (a kind gift from Scott Lowe). *Tnfaip3*<sup>-/-</sup> MEFs were a kind gift from Averil Ma. Primary and immortalized MEFs were cultured in FMA medium (DMEM supplemented with 10% heat inactivated fetal bovine serum, 100 µM non-essential amino acids (Invitrogen), 50 µM 2-mercaptoethanol, 1% penicillin/streptomycin, and 1% L-glutamine). A20 macrophage progenitor cells were generated by isolating bone marrow cells from the femurs and tibias of 6–8-week-old mice, followed by erythrocyte lysis and enrichment using a murine haematopoietic progenitor cell enrichment kit as directed by the manufacturer (19756; Stem Cell Technologies). Cells were then immortalized using conditional HoxB8. Progenitor cells were cultured in RPMI-1640 medium supplemented with 10% heat-inactivated fetal bovine serum, 2 mM L-glutamine, 20 ng ml<sup>-1</sup> GM-CSF and 1 µM β-oestradiol. Differentiation to macrophages was by removal of β-oestradiol. Primary BMDMs were generated by isolating the bone marrow cells from mice as described above and were cultured in FMA media supplemented with 20 ng ml<sup>-1</sup> M-CSF for 7 days. Primary and immortalized cells were characterized by genotyping and western blotting as described and tested for mycoplasma contamination. HEK293T cells were authenticated following Genentech's "Guidelines for Maintaining the Integrity of Cell Line Stocks" as described previously<sup>49</sup> and were cultured in DMEM supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin, and 1% L-glutamine.

**In vitro cytokine activation treatments.** For signalling pathway profiling studies MEFs or BMDM (primary or matured from HoxB8-ER immortalized progenitors) were treated with 20 ng ml<sup>-1</sup> human TNF or 100 ng ml<sup>-1</sup> LPS for 30–45 min to induce A20 expression, washed three times with PBS, and cultured in FMA or RPMI medium until collection at the indicated time points. Alternatively primary BMDMs and MEFs were treated with 20 ng ml<sup>-1</sup> to 1 µg ml<sup>-1</sup> LPS or human TNF (with or without the Flag epitope fusion, as indicated) until collection at the indicated time points. For analysis of signalling complexes cells were treated with 20 µM MG-132 immediately before collection (untreated controls) or immediately before treatment with TNF or Flag-TNF as indicated, or 10 µg ml<sup>-1</sup> LPS, and collected at the indicated time points.

**Western blot analysis and immunoprecipitations.** Cells were treated as described and lysed TNFR1 lysis buffer (20 mM Tris pH 7.5, 150 mM NaCl, 1% Triton X-100, 1 mM EDTA, 50 mM NaF, 10 mM *N*-ethyl maleimide, complete protease inhibitor tablets (Roche), phosphatase inhibitor cocktail-1 and -3 (Sigma), and 25 µM MG-132) containing 6 M urea. Lysates were reduced and alkylated and processed as previously described<sup>8</sup>. To evaluate TRAF6 ubiquitination status, cells were first treated as detailed above and lysed in LPS lysis buffer (20 mM HEPES pH 7.6, 150 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 2 mM EDTA, 0.5% Triton X-100, 10 mM NaF, 2 mM



DTT, 10 mM *N*-ethyl maleimide, complete protease inhibitor tablets (Roche), Phosphatase inhibitor cocktail-1 and -3 (Sigma), and 25  $\mu$ M MG-132) containing 6 M urea at the indicated time points. Protein concentrations were quantified and 200  $\mu$ g was reserved for western blot analysis. Normalized lysates from each sample were subsequently diluted to 4 M urea with LPS IP buffer and samples were pre-cleared with protein A beads and 2  $\mu$ g per mg total protein non-production grade Trastuzumab. Pre-cleared lysates were subsequently immunoprecipitated with 1  $\mu$ g anti-K63 antibody per mg total protein overnight, immunocomplexes were captured with 10  $\mu$ l Protein A beads (Sigma) per mg total protein, washed with LPS IP buffer, and processed for western blot analysis. To evaluate RIPK1 or TNFR1 ubiquitination status, cells were first treated as detailed above using TNF or Flag-TNF and lysed in TNFR1 lysis buffer at the indicated time points. Protein concentrations were quantified and 200  $\mu$ g was reserved for western blot analysis. Normalized lysates from each sample were pre-cleared with protein A+G beads (Pierce) or with anti-IgG beads (Sigma). Pre-cleared lysates were immunoprecipitated with 5  $\mu$ g anti-murine TNFR1 antibody pre-coupled to 50  $\mu$ l protein A+G beads or with anti-Flag beads (Sigma), washed, and the immunoprecipitates were dissociated from the beads with 6 M urea. The eluted proteins were subsequently diluted to 4 M urea for anti-K63 ubiquitin IPs, or remained undiluted for anti-linear ubiquitin IPs, in ubiquitin chain lysis buffer (20 mM Tris-Cl pH 7.5, 135 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 1% Triton X-100, 1 mM EGTA, 10% glycerol, 50 mM NaF, 10 mM *N*-ethyl maleimide, complete protease inhibitor tablets (Roche), Phosphatase inhibitor cocktail-1 and -3 (Sigma), and 25  $\mu$ M MG-132) and pre-cleared with protein A beads and 2  $\mu$ g per mg total protein non-production grade Trastuzumab, followed by immunoprecipitation with 1  $\mu$ g anti-K63 or anti-linear antibody per mg total protein overnight, and immunocomplexes were captured with 10  $\mu$ l Protein A beads (Sigma) per mg total protein. Immunoprecipitates were then washed and processed for western blot analysis. For analysis of Flag-TNF activated TNFR1 signalling complexes, cells were first treated as described above for the indicated time points. Cells were immediately washed with PBS and lysed at 4°C in TNFR1 lysis buffer and in some cases stored at -80°C. Lysates were cleared by centrifugation, pre-cleared with mouse IgG agarose (Sigma), and normalized amounts of lysates were immunoprecipitated with anti-Flag affinity gel (Sigma) overnight. Immunoprecipitates were washed once with wash buffer #1 (20 mM HEPES pH 7.9, 420 mM NaCl, 1.5 mM MgCl<sub>2</sub>, 0.2 mM EDTA, 25% glycerol, complete protease inhibitor cocktail) and four times with wash buffer #2 (20 mM Tris pH 7.4, 20% glycerol, 0.2 mM EDTA, 300 mM NaCl, 0.1% NP-40, complete protease inhibitor cocktail), rotating at least 10 min for each wash in buffer #2. Samples were eluted with 500  $\mu$ g per ml 3 $\times$ Flag peptide (Sigma), concentrated, and prepared for western blot analysis. For analysis of FADD-associated signalling complexes, cells were treated with RNAi oligonucleotides as outlined in the 'RNAi treatments and transfections' section and treated with TNF as outlined above. Cells were immediately washed with PBS and lysed at 4°C in TNFR1 lysis buffer as described above containing 10  $\mu$ M Z-VAD (R&D). Lysates were cleared by centrifugation, protein concentrations were quantified and 200  $\mu$ g was reserved for western blot analysis. Normalized amounts of lysates were pre-cleared with Protein G beads (Sigma) and immunoprecipitated with anti-FADD antibody (Genentech) overnight. Immunocomplexes were captured with Protein G affinity gel matrix (Sigma), washed with TNFR1 IP buffer, and processed for western blot analysis. For analysis of K63-polyubiquitinated RIPK1 associated with FADD, cells were treated and lysates were prepared and immunoprecipitated with anti-FADD antibody and Protein G affinity gel matrix as outlined above. The washed immunoprecipitates were dissociated from the beads with 6 M urea. The eluted proteins were subsequently diluted to 4 M urea in ubiquitin chain lysis buffer and pre-cleared with protein A beads and 2  $\mu$ g per mg total protein non-production grade Trastuzumab, followed by immunoprecipitation with 1  $\mu$ g anti-K63 antibody per mg total protein overnight, and immunocomplexes were captured with 10  $\mu$ l Protein A beads (Sigma) per mg total protein. Immunoprecipitates were then washed and processed for western blot analysis. For immunoprecipitations using anti-linkage-specific ubiquitin antibodies, cells were treated as indicated and lysed in ubiquitin lysis buffer with 6 M urea. Normalized lysates were either undiluted (for anti-linear ubiquitin immunoprecipitations) or diluted to 4 M urea (for anti-K11, anti-K48, or anti-K63 ubiquitin immunoprecipitations). Anti-K63 immunoprecipitations were performed as described above, 2  $\mu$ g anti-K11 antibody per mg of total protein was used for anti-K11 immunoprecipitations, and 5  $\mu$ g anti-K48 ubiquitin antibody (Genentech) + 1  $\mu$ g anti-K48 ubiquitin antibody (CST) per mg total protein were used for anti-K48 immunoprecipitations. Samples were incubated with antibodies overnight at 4°C and captured with protein A beads and processed as above.

**Induction of myelin oligodendrocyte glycoprotein (MOG) peptide 35-55 EAE.** To select animals for the study, age and gender matching between genotypes was performed. Animals were used based on their genotype information and thus

could not be randomized. Sample size was generated based on historical data and knowledge of variability within the EAE model.

EAE was induced in 8–12 week old female wild-type A20, A20 OTU(C103A), or A20 ZnF4(C609A,C612A) mice. Briefly, animals were injected subcutaneously on the back with a 200  $\mu$ l emulsification of 300  $\mu$ g MOG peptide in incomplete Freund's adjuvant supplemented with 8 mg ml<sup>-1</sup> Mycobacterium tuberculosis (Difco Laboratories, Detroit, MI). The final dose of Mycobacterium tuberculosis per mouse was 800  $\mu$ g. Clinical scoring of disease was performed 3 times per week starting at day 9. Cages were marked with genotype information that, while not referenced during scoring, were present and not blinded to the investigator.

Animals were assessed based on a five point system: 0 = no clinical disease, 1 = loss of tail tone only, 2 = mild monoparesis or paraparesis, 3 = severe paraparesis, 4 = paraplegia and/or quadraparesis, and 5 = moribund or death. Average daily clinical score (ADCS) was calculated as the area under the curve (AUC) value for clinical score from day of disease onset to end of study divided by the duration in days. Animals were identified by numbers and the investigators scoring the animals were blinded to the genotype/number correlations.

**EAE specimen preparation and histopathology analysis.** Spines and brain were collected at terminal euthanasia and fixed in 10% neutral buffered formalin. Spines were decalcified until trimmable. CNS was evaluated in 4 coronal sections (brain) and 5 to 8 transverse sections (spinal cord) stained with haematoxylin, eosin and Luxol fast blue. Lesion severity was scored on an arbitrary scale of 0 to 3 and reported as averages of all sections scored per animal.

**Mass spectrometric identification of ubiquitinated TNFR1.** Profiling of ubiquitination sites were performed using PTMscan protocol (Cell Signaling Technology, Danvers, MA). Briefly, wild-type and OTU mutant MEF cells were treated with TNF $\alpha$  for 15 min and immediately stopped by adding 9 M urea lysis buffer (20 mM HEPES, pH 8.0, 9 M urea, 1 mM sodium vanadate, 2.5 mM sodium pyrophosphate, 1 mM beta-glycerophosphate). Lysates were reduced with dithiothreitol, alkylated with iodoacetamide and digested with trypsin or chymotrypsin+trypsin combo at room temperature for overnight. Resultant peptides were desalted on Sep-pak C18 cartridges (Waters, Milford, MA) and lyophilized for two days. Immunoprecipitation of ubiquitinated peptides was performed using anti-K- $\epsilon$ -GG antibody (Cell Signaling Technology, Danvers, MA). Peptides were eluted with 0.15% TFA, desalted, and further separated into 5 fractions with high pH reverse phase on a STAGE tip. All peptides were analysed with a NanoAcquity UPLC system (Waters, Milford, MA) directly coupled to an LTQ Orbitrap Elite mass spectrometer (Thermo Scientific, San Jose, CA). Peptides were reconstituted in 0.1% formic acid (FA) with 2% acetonitrile (ACN), loaded onto a Symmetry C18 column (1.7 mm BEH-130, 0.1  $\times$  100 mm, Waters) and separated with a 60-min gradient from 0% to 15%, 0% to 20%, or 2% to 25% solvent B (0.1% FA, 98% ACN) at 1  $\mu$ l min<sup>-1</sup> flow rate. Peptides were eluted directly into the mass spectrometer with a spray voltage of 1.2 kV. Full MS data were acquired in FT for 375–1,600 *m/z* with a 60,000-resolution. The 15 most abundant ions found in the full MS were selected for MS/MS through a 2-Da isolation window.

Acquired MS/MS spectra were searched using the Mascot (Matrix Sciences, London, UK) with trypsin or trypsin+chymotrypsin enzyme specificity. Search criteria included a full MS tolerance of 50 ppm, MS/MS tolerance of 0.8 Da with oxidation (+15.9949 Da) of methionine and ubiquitination (+114.0429 Da) of lysine as variable modifications and carbamidomethylation (+57.0215 Da) of cysteine as static modifications. Data were searched against the mouse and contaminant subset of the Uniprot database that consists of the reverse protein sequences. Identified TNF-R1 ubiquitinated peptide-spectrum-matchings (PSMs) were manually validated. Additional isotopically-labelled peptides (AQUA) corresponding to the identified TNF-R1 ubiquitinated sequences were synthesized by Cell Signaling Technologies (Danvers, MA) to confirm the endogenous ubiquitination. Abundance of ubiquitinated peptides was determined by adding equal amount of AQUA peptides to the wild type and mutant MEF samples before immunoprecipitation, followed by quantifying area under curve for the corresponding spiked-in and endogenous peptides.

**In vitro deubiquitinase reactions.** Recombinant full length A20, with or without the indicated amino acid mutations, was expressed in *E. coli* and purified as described previously<sup>8,9</sup>. Recombinant Flag-tagged full length A20, with or without the indicated amino acid mutations, was expressed in HEK-293T cells and purified as described previously<sup>9</sup>. The input of all A20 proteins for *in vitro* deubiquitinase reactions was normalized by estimation of the protein concentrations on Coomassie blue-stained gels using serial dilutions of bovine serum albumin as standards. *E. coli*-derived A20 was phosphorylated with recombinant IkK $\beta$  (Prokinase) in the following reaction: 5  $\mu$ g A20, 1.75  $\mu$ g GST-IkK $\beta$ , 10  $\mu$ M ATP, 25 mM Tris pH 7.5, 5 mM  $\beta$ -glycerolphosphate, 1 mM DTT, 0.1 mM Na<sub>3</sub>VO<sub>4</sub>, 10 mM MgCl<sub>2</sub>, 0.5% phosphatase inhibitor cocktail-3. A20 deubiquitination reactions were performed using 100 ng recombinant A20, 500 ng of the indicated



ubiquitin chain (unconjugated or conjugated to HA epitope-tagged RIPK1 peptide, see below), and DUB reaction buffer (50 mM HEPES pH 8.0, 0.01% Brij-35, and 3 mM DTT) with or without phosphatase inhibitor cocktail and were incubated for the indicated times at 37 °C with agitation at 1,000 rpm. Ubiquitinated TNFR1 or RIPK1 substrate was purified from Flag–TNF-treated *Tnfr1*<sup>3<sup>OTU</sup></sup> MEFs as outlined above. OTULIN deubiquitinase reactions were performed following the recommended UbiCREST protocol conditions (Boston Biochem). Samples were subsequently processed for western blot analysis as described above and immunoblotted as indicated.

**Preparation of polyubiquitinated RIPK1 peptide.** The sequence of the HA–RIPK1 peptide is as follows:

YPYDVPDYASLEHPQEENPSLQSKLQDEANYHLYGSRMDRQT-amide.

The peptide was prepared at Genentech on a Protein Technologies Symphony automated synthesizer using standard Fmoc chemistry protocols on a Fmoc Rink amide linker attached to TentGel resin. Peptides were cleaved off the solid support with trifluoroacetic acid: triisopropylsilane: water (95:2.5:2.5) for 2 h at room temperature. Trifluoroacetic acid was evaporated and peptides were precipitated with ethyl ether, extracted with acetic acid, acetonitrile and water and lyophilized. Crude peptides were solubilized in dimethyl sulfoxide and purified by reverse phase chromatography on a C18 column using acetonitrile/water buffers. Purified fractions were analysed by liquid chromatography mass spectrometry (PE–Sciex), pooled and lyophilized. The calculated and the found mass were both 5111.3 Da. K48-linked or K63-linked tetraubiquitin chains were then ligated to the peptide using the following reaction mix: 30 μM tetraubiquitin, 0.5 μM E1 (Boston Biochem), 5 μM cdc34 (for K48 tetraubiquitin, Genentech) or 2.5 μM each UEV1 and UbcH13 (for K63 tetraubiquitin, Genentech), 50 mM Tris pH 8.0 (Genentech), 10 mM ATP (Sigma), 10 mM MgCl<sub>2</sub> (Genentech), and 6 mM DTT (Sigma). Reactions were incubated with agitation for 4 h at 30 °C and then overnight at 27.5 °C. Tetraubiquitin-conjugated peptides were purified and concentrated using 30 kDa cutoff microspin columns (Amicon Ultra). K63-linked tetraubiquitin chains ligated to the HA-peptide were subsequently modified with linear ubiquitin chains using 30 μM linear tetraubiquitin (Genentech), 5 μM UbcH5c (Boston Biochem), 1 μM His6-MBP HOIP catalytic domain (Boston Biochem), 50 mM Tris pH 8.0 (Genentech), 10 mM ATP (Sigma), 10 mM MgCl<sub>2</sub> (Genentech), and 6 mM DTT (Sigma). Reactions were incubated with agitation for 4 h at 30 °C and then overnight at 27.5 °C. Polyubiquitin-conjugated peptides were purified and concentrated using 30 kDa cutoff microspin columns (Amicon Ultra).

**Identification of A20 phosphorylation sites by mass spectrometric analysis.** Human and murine A20 proteins were reduced in sample buffer containing DTT (Sigma, St Louis, MO) and alkylated in 0.176 M *n*-isopropylidiodoacetamide (synthesized in house, mass addition of 99.0684 Da to all cysteine residues). Samples were separated by SDS–PAGE and stained. Bands around 90 kDa in each lane were excised and digested with trypsin as previously described<sup>50</sup>. After overnight digestion the peptides were extracted from gel slices in acetonitrile and evaporated to near dryness. Samples were then reconstituted in 10 μl of 0.1% formic acid containing 2% acetonitrile and analysed by LC–MS/MS. Reconstituted peptides were injected via an auto-sampler onto a 75 μm × 100 mm column (BEH, 1.7 μm, Waters Corp, Milford, MA) at a flow rate of 1 μl min<sup>−1</sup> using a NanoAcquity UPLC (Waters Corp, Milford, MA). A gradient from 98% Solvent A (water + 0.1% formic acid) to 80% Solvent B (acetonitrile + 0.08% formic acid) was applied over 45 min. Samples were analysed on-line via nanospray ionization into a hybrid LTQ–Orbitrap Velos mass spectrometer (Thermo, San Jose, CA). Data were collected in data dependent mode with the parent ion analysed in the FTMS and the top 15 most abundant ions selected for fragmentation in the LTQ. Tandem mass spectrometric data were analysed using the search algorithm Mascot (Matrix Sciences, London, UK). Searches were performed against the UniProt database with a parent ion tolerance less than 50 ppm, fixed carbamylation (NIPCAM on Cys) and variable oxidation (Met) and phosphorylation (Ser, Thr, or Tyr). Phosphorylation sites were localized by *de novo* interpretation of the spectra and using Ascore (Harvard University, Cambridge, MA) as previously described<sup>50</sup>.

**RNAi treatments and transfections.** HEK293T cells were transfected with three On Target Plus TNFRSFA1 siRNA SMARTpool (Thermo Scientific) oligonucleotides using RNAi Max transfection reagent (Life Technologies) according to the manufacturer's instructions. The RNAi sequences are:

Oligo 1: sense = CAAAGGAACCUACUUGUACUU.

Oligo 2: sense = GAGCUUGAAGGAACUACUAAU.

Oligo 4: sense = UCCAAGCUCUACUCCAUGUU.

MEFs were transfected with On Target Plus RNF31 siRNA SMARTpool (Thermo Scientific) using RNAi Max transfection reagent (Life Technologies) according to the manufacturer's instructions. The RNAi sequences are:

Oligo 1: sense = GCUGCAAGGUGCCGGGAAU.

Oligo 2: sense = GCUAAGAGAGAGCGUUGAA.

Oligo 3: sense = GCCAAGAUAGAUGCGGAA.

Oligo 4: sense = GGCAUUGACUGUCCGAAAU.

To validate sufficient knockdown, lysates were probed using anti-HOIP antibodies and/or transcript levels were evaluated for HOIP expression. To this end, RNA was isolated from transfected cells and Taqman probes (Life Technologies) were used in a 1-step RT–PCR reaction (HOIP: Mm01313902\_m1 and GAPDH: Mm99999915\_g1). Relative quantification of transcript was determined by comparing normalized C<sub>t</sub> values.

To introduce a K256R point mutation in wild-type murine TNFR1 (mTNFR1) the following primers were used in a QuikChange mutagenesis reaction:

mTNFRK256R forward primer: TGT AGG GAT CCC GTG CCT GTC AGA GAG GAG AAG GCT GGA AAG.

mTNFRK256R reverse primer: CTT TCC AGC CTT CTC CTC TCT GAC AGG CAC GGG ATC CCT ACA.

The resulting constructs were then sequenced to confirm the successful mutagenesis and a second PCR reaction was carried out using the primers below to add and an HA tag on the N terminus of both wild-type and mutated mTNFR1.

mTNFR XbaI reverse primer: CGG TCT AGA TTA TCG CGG GAG GCG GGT CGT.

The resulting PCR products and pCDNA3.1 vector were digested with EcoRI and XbaI restriction enzymes and then ligated together. Plasmids were sequenced to confirm the presence of the HA tag.

**Live cell imaging experiments.** For measurement of cell viability and caspase activity, 2,500 MEFs that were previously transfected with the indicated siRNA oligonucleotides were seeded per well of a 96-well plate (Corning catalogue # 3904) in media containing 2 μM CellEvent Caspase 3/7 reagent (Life Technologies catalogue # C10423). The following day, the cells were treated with TNF and immediately placed in an IncuCyte live cell imager. Images were taken every 2 h using a 10× objective. All treatments were done in triplicate. Phase contrast was used to measure cell confluency/density while green fluorescence was used to measure caspase activity. The images were analysed using IncuCyte software (Basic Analysis parameters) and a ratio of caspase activity to cell density was determined. The area under the curve (AUC) for caspase activity/cell density was determined for each treatment. Student *t*-tests were used to measure statistical significance and error bars were calculated to indicate s.e.m.. To measure cell death, 0.1 μM ethidium homodimer-1 (Life Technologies catalogue # E1169) was added to the media before treatment with TNF. Cells were monitored using IncuCyte software as indicated above. Dead cells were counted based on the intensity of red fluorescence. Red counts were normalized to cell density. The area under the curve (AUC) for dead cells/cell density was determined for each treatment. Student *t*-tests were used to measure statistical significance and error bars were calculated to indicate s.e.m.

**Identification of Flag–TNF-associated proteins.** Cells were treated with recombinant Flag–TNF for the indicated times and samples prepared as described in the 'Western blot analysis and immunoprecipitations' section. The TNF receptor complex was immunoprecipitated through anti-Flag antibody conjugated agarose gel (Sigma). Proteins were eluted off beads using 3 × Flag peptide (Sigma) and concentrated with a 10K cutoff membrane filter. Eluents were mixed with LDS sample buffer (Invitrogen), reduced with dithiothreitol (DTT), alkylated with Iodoacetamide and resolved by a 3–8% Tris acetate gel. Each lane was cut into 5 bands and subjected to in-gel digestion. Briefly, each gel band was digested with trypsin overnight at 37 °C in the presence of 25 mM AMBIC at pH 8.0 for extraction. Peptides were further extracted by 10% acetonitrile with 0.1% trifluoroacetic acid. Extracts were combined and dried under vacuum. Samples were reconstituted in 2% acetonitrile with 0.1% formic acid and loaded onto a 0.1 × 100 mm analytical column packed with 1.7 μm BEH-130 C18 using a NanoAcquity UPLC (Waters). Peptides were eluted with a 60 min gradient from 2% to 25% acetonitrile at 1 μl min<sup>−1</sup> flow rate and directly introduced to an LTQ–Orbitrap Elite mass spectrometer (ThermoFisher Scientific) through an ADVANCE electrospray ionization source (Michrom BioResources/Bruker, Auburn, CA). Full MS data were acquired in orbitrap at 60,000 resolutions. The top 15 most abundant precursors from the preceding full MS spectrum were further selected for CID fragmentation and MS/MS spectra were acquired in ion trap. MS/MS data was searched using the Mascot (version 2.3, Matrix Sciences, London, UK). Search criteria included a full MS tolerance of 50 ppm, MS/MS tolerance of 0.8 Da with oxidation (+15.9949 Da) of methionine and ubiquitination (+114.0429 Da) of lysine as variable modifications and carbamidomethylation (+57.0215 Da) of cysteine as static modifications with up to 3 missed cleavages. Data was searched against the *Mus musculus* and contaminant subset of the Uniprot database that consists of the reverse protein sequences. Data was then filtered using linear discriminator analysis (LDA) at peptide level to 10% false discovery rate (FDR). Further cutoff was applied to the whole data set at the protein level to 5% FDR, which resulted in 0.6% FDR at peptide level.

**Cloning, expression and purification of A20 ZnF proteins.** The human A20 ZnF1 (K386–S453), ZnF4 (S592–K635), ZnF7 (P758–G790) domains and mutants (ZnF4(C624A,C627A); ZnF4(K606E,I629R); ZnF4(I629R); ZnF4(K606E); and ZnF7(F770A,G771A)) were cloned into the EitNTH-NAvi vector with N-terminal His6 tag, Avi tag and TEV cleavage site and further transformed into BL21-Gold (DE3) *E. coli* strain. Expression of  $^{15}\text{N}$ -labelled proteins for NMR studies was carried out in M9 media at 16 °C for approximately 20 h using 0.4 mM IPTG induction. All biotinylated proteins were expressed in the same *E. coli* strain. The purification of all proteins was carried out at 4 °C using Ni-NTA resin (Qiagen) followed by protease cleavage and another  $\text{Ni}^{2+}$  affinity chromatography to remove the purification tag. Proteins were further purified by size exclusion chromatography (Superdex 75) using buffer consisting of 20 mM Tris (pH 8.3), 300 mM NaCl, and 1 mM Tris (2-carboxyethyl) phosphine (TCEP).

**NMR experiments and data analysis.** NMR experiments were performed at 25 °C on a Bruker 500 MHz spectrometer. All NMR samples were prepared in buffer containing 20 mM MES (pH 6.0), 150 mM NaCl, 0.5 mM TCEP and 10% (v/v)  $\text{D}_2\text{O}$  with concentration 0.1–0.13 mM of A20 ZnF motifs. Mono-ubiquitin was added to the A20 ZnF motifs at 0 to 10 ratios using a 11.1 mM stock of mono-ubiquitin. The data were analysed and  $K_D$  values were determined using Sparky and NMRViewJ software, respectively.

***In vitro* A20 zinc finger ubiquitin-binding assays.** Binding of A20 ZnF motifs to Ub trimers of varying linkages was measured by biolayer interferometry on an Octet RED384 instrument (ForteBio, Inc.). Biotinylated ZnF1, ZnF4 and ZnF7 variants were captured onto streptavidin SA biosensors. Unbound material was washed away with binding buffer (20 mM MES pH 6.0, 150 mM NaCl, 10% glycerol, 0.2 mM DTT, 0.01% Tween-20 and 0.1  $\text{mg ml}^{-1}$  human serum albumin) before conducting association and dissociation measurements with trimeric Ub analytes. The binding reactions displayed rapid saturation behaviour and thus were ideally evaluated by equilibrium binding analysis; however, because trimeric ubiquitin chains possess multiple binding sites for the ZnF motifs, these measurements were encumbered by surface-dependent avidity. These binding curves displayed multiphasic behaviour — the true binding phases were complicated by additional, artificially high affinity phases, in which adjacent ZnF molecules affixed to the tip surface engaged the same ubiquitin trimer in an avid interaction. In order to avoid these artefacts, we identified conditions in which avid interactions were minimized as follows: using linear triubiquitin binding ZnF7 as a test case, we performed titrations at a range of ZnF7 loading densities, plotted equilibrium response values (in nm) as a function of linear ubiquitin trimer concentration, and fit the curves to a modified two-site binding equation:

$$R = R_{\max} \times \left\{ \left[ (1 - n_{\text{avid}}) / (K_D + [A]) \right] + \left[ n_{\text{avid}} / (K_{D-\text{avid}} + [A]) \right] \right\}$$

in which  $R$  is the response value (in nm),  $R_{\max}$  is the maximal response,  $n_{\text{avid}}$  is the relative fraction of avid interactions,  $[A]$  is the total concentration of Ub trimer,  $K_D$  is the equilibrium dissociation constant for the non-avid interactions, and  $K_{D-\text{avid}}$

is the apparent equilibrium dissociation constant for the avid interactions. This equation reflects the assumptions that avid interactions have a much stronger affinity than non-avid interactions, and the stochastic placement of ZnF7 molecules on the tip surface ensures that some non-avid interactions will occur regardless of loading density. We plotted the resulting  $n_{\text{avid}}$  values as a function of ZnF7 loading density to determine the ZnF loading range in which the fraction of avid interactions reached a minimal plateau. This occurred at a loading density at or below 0.12 nm, and so all subsequent measurements were conducted using this loading density. Once avid interactions were minimized, the equilibrium response data, in most cases, was well fit by a simple, 1:1 binding equation:

$$R = R_{\max} \times [A] / (K_D + [A]),$$

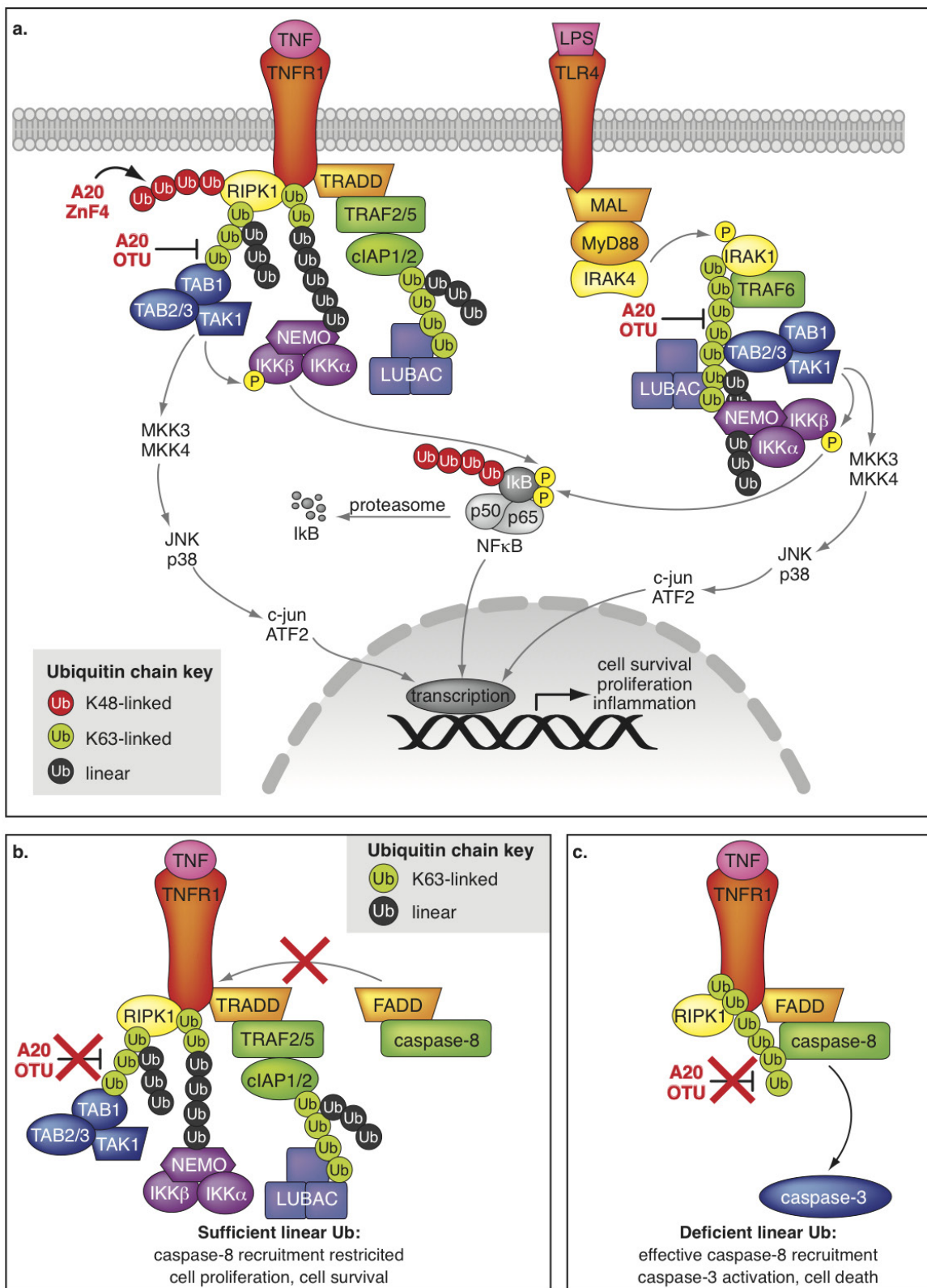
where  $R$  is the response value,  $R_{\max}$  is the maximal response,  $[A]$  is the total Ub trimer analyte concentration, and  $K_D$  is the equilibrium dissociation constant. The exception was ZnF7 binding to linear Ub trimer. Because ZnF7 has two binding sites for ubiquitin, these data were fit to a standard 2-site binding model:

$$R = [R_{\max 1} \times [A] / (K_{D1} + [A])] + [R_{\max 2} \times [A] / (K_{D2} + [A])]$$

Here each of the two equilibrium constants ( $K_{D1}$  and  $K_{D2}$ ) has a separate  $R_{\max}$  value ( $R_{\max 1}$  and  $R_{\max 2}$ , respectively). All data were fit using Kaleidagraph version 4.03 (Synergy Software).

***In vitro* ubiquitination assays.** *E. coli*-derived A20 was phosphorylated with recombinant  $\text{I}\kappa\text{K}\beta$  (Prokinase, Active Motif or Life Technologies) in the following reaction: up to 5  $\mu\text{g}$  A20, up to 1.75  $\mu\text{g}$  GST- $\text{I}\kappa\text{K}\beta$ , 10  $\mu\text{M}$  ATP, 25 mM Tris pH 7.5, 5 mM  $\beta$ -glycerolphosphate, 1 mM DTT, 0.1 mM  $\text{Na}_3\text{VO}_4$ , 10 mM  $\text{MgCl}_2$ , 0.5% phosphatase inhibitor cocktail-3. A20 ubiquitination reactions were performed as previously described<sup>8</sup> using UbcH5a or UbcH7 as E2 enzymes (Boston Biochem), Flag-wild-type A20 or Flag-A20 ZnF4(C609A,C612A) expressed in HEK293T cells and purified by Flag peptide elution, or *E. coli*-derived A20 with or without  $\text{I}\kappa\text{K}\beta$  phosphorylation as described above were used as ubiquitin ligases, and the HA-RIPK1 peptide or recombinant murine TNFR1 (Genentech) were used as substrates. For assessment of TNFR1 K48-ubiquitination *in vitro*, after ubiquitination reactions were complete 6 M urea was added to each reaction for 15 min at room temperature with agitation to dissociate proteins. The urea was diluted to 4 M and the reactions were immunoprecipitated with 5  $\mu\text{g}$  anti-K48 ubiquitin antibody (Genentech) + 1  $\mu\text{g}$  anti K48 ubiquitin antibody (CST) overnight at 4 °C. The immunocomplexes were captured with protein A beads and processed as above for western blot analysis.

49. Yu, M. *et al.* A resource for cell line authentication, annotation and quality control. *Nature* **520**, 307–311 (2015).
50. Wertz, I. E. *et al.* Sensitivity to antitubulin chemotherapeutics is regulated by MCL1 and FBW7. *Nature* **471**, 110–114 (2011).



Extended Data Figure 1 | See next page for caption.



**Extended Data Figure 1 | A model for A20 OTU and A20 ZnF4 regulation of TNF- and LPS-activated signalling.**

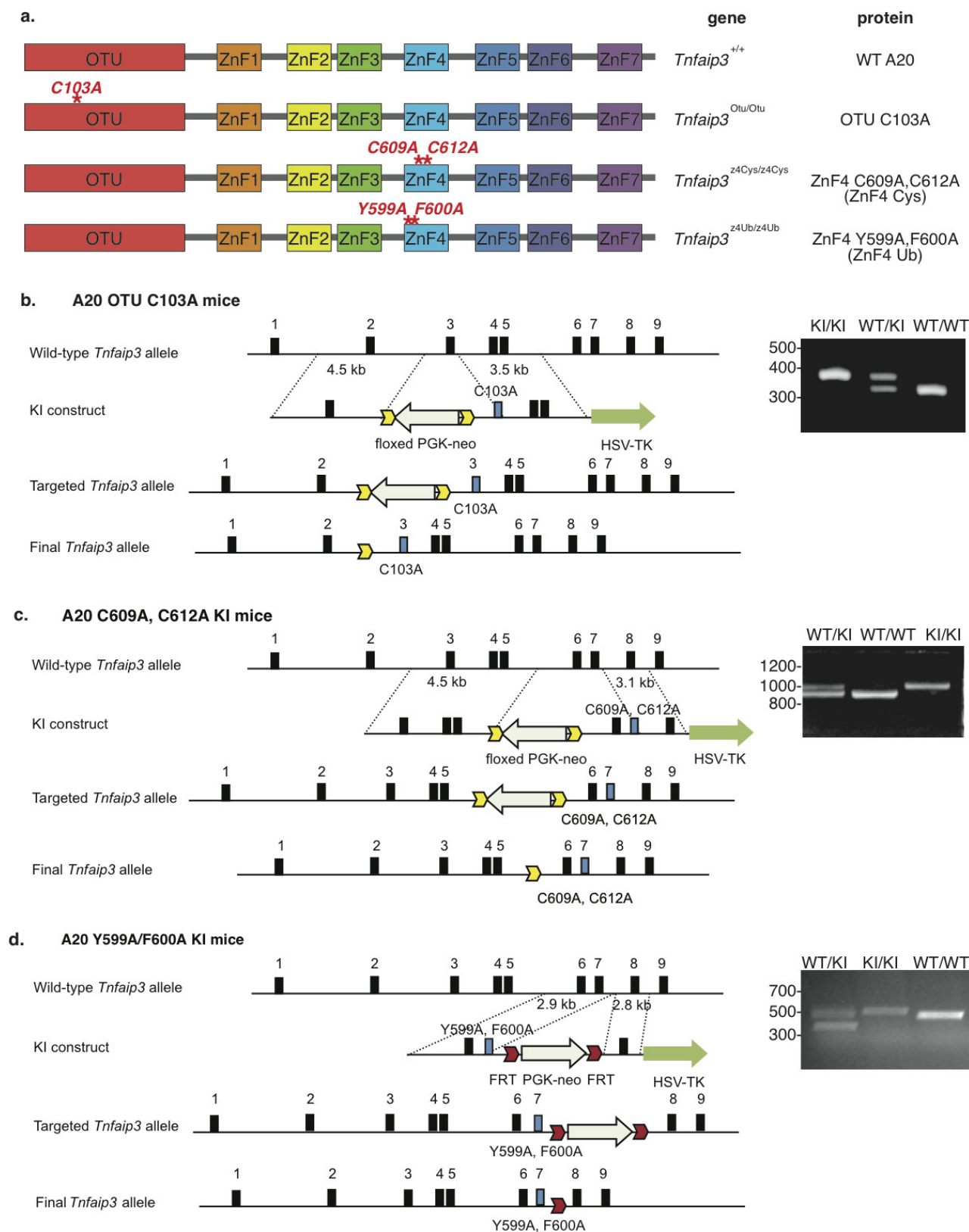
**a, Left complex.** Upon TNF binding TNFR1 forms a trimer, thereby promoting recruitment of the adaptor protein TRADD and the RIP1 kinase (RIPK1). TRADD recruits TRAF2, TRAF5 and the ubiquitin ligases cIAP1 and cIAP2. The cIAP proteins promote K63-linked ubiquitination of signalling proteins including RIPK1, cIAP1/2 (autoubiquitination), and possibly TNFR1. K63 ubiquitination of cIAP1/2 subsequently recruits the LUBAC complex that promotes linear polyubiquitination of signalling proteins including RIPK1 and TNFR1. K63 ubiquitin chains on RIPK1 promote recruitment of the TAK1/TAB2/3 complex, whereas linear ubiquitin chains on RIPK1 promote I $\kappa$ B kinase complex recruitment via NEMO. Kinase complex recruitment promotes their subsequent activation and propagation of downstream JNK, p38 (via MKK3 and MKK4) and NF $\kappa$ B signalling pathways. We propose that A20 is recruited to the active TNFR1 signalling complex via ZnF7 binding to linear ubiquitin chains. The A20 OTU domain catalytic C103 is essential for attenuating TNF-activated signalling by removing K63 polyubiquitin chains from RIPK1 and other proteins including TNFR1, thereby promoting the dissociation of the active signalling complex. The A20 ZnF4 motif, that depends on C609/C612 for structural integrity and Y599/F600 for ubiquitin binding, is likely to collaborate with other proteins (not shown) to further downregulate TNF signalling by directing K48 polyubiquitination and subsequent degradation of proximal complex proteins, including RIPK1 and TNFR1.

**Right complex.** LPS binding activates TLR4 and promotes the assembly of proximal signalling complexes via the adaptors TRIF and TRAM

(not shown) or Mal and MyD88. Recruitment and activation of the proximal kinases IRAK4 and IRAK1, the ubiquitin ligase Pellino (not shown), and the LUBAC complex promote K63 and linear polyubiquitination of signalling proteins. As with TNFR1 signalling, this scaffolding-type ubiquitination promotes recruitment of TAK1/TAB2/3 and I $\kappa$ B kinase complexes, their subsequent activation, and propagation of downstream JNK, p38 and NF $\kappa$ B signalling pathways. A20 is probably recruited to the LPS-activated signalling complex via ZnF7 binding to linear ubiquitin chains. The A20 OTU domain catalytic C103 is essential for attenuating LPS-activated signalling by removing K63 polyubiquitin chains from TRAF6, and possibly other proximal signalling proteins. Although the structural integrity and the ubiquitin-binding function of A20 ZnF4 is dispensable for proper attenuation of TLR4 signalling, A20 ZnF4 could have a redundant function with another protein.

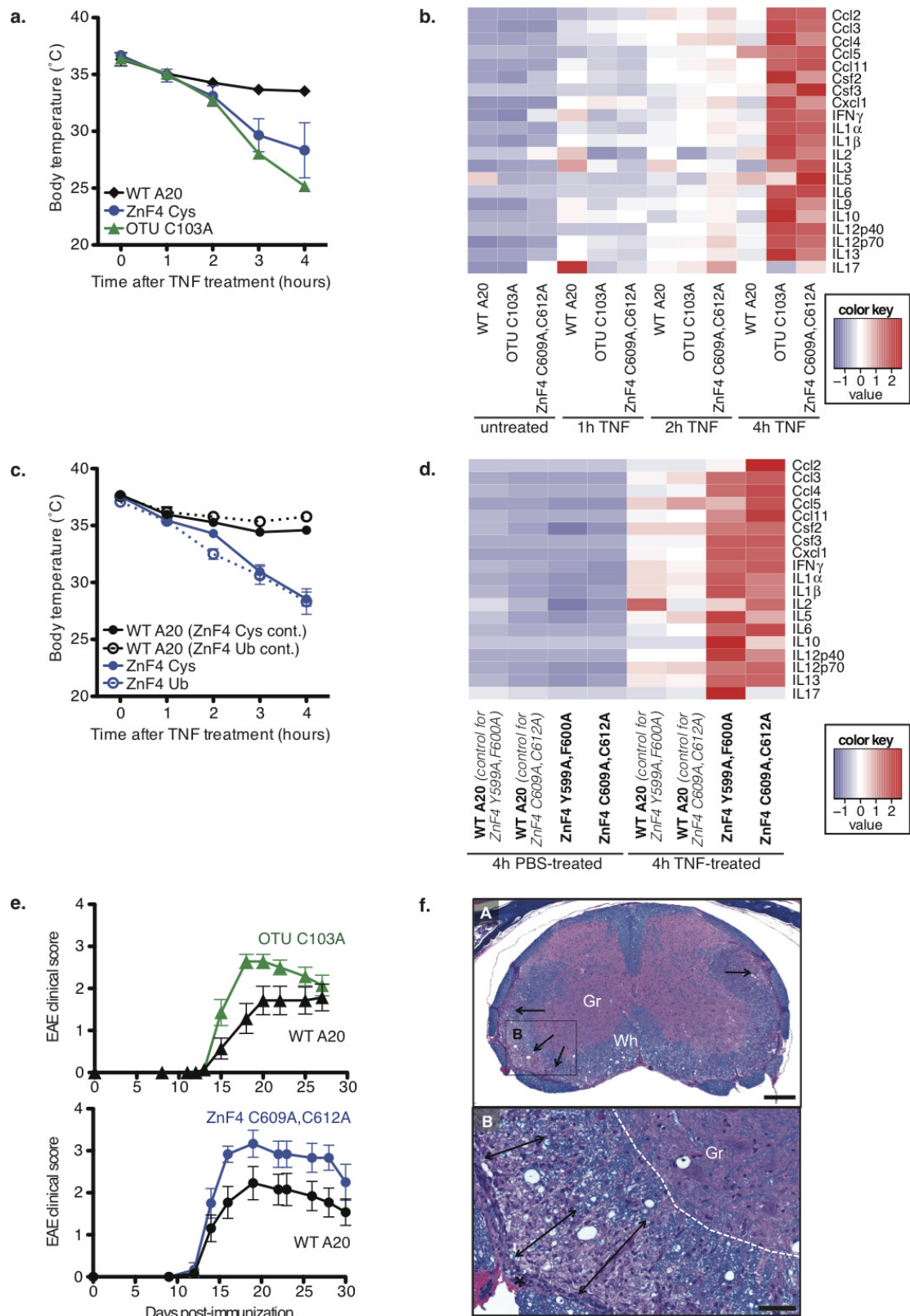
**b, In A20 OTU(C103A) cells** removal of K63 ubiquitin chains on proximal signalling components is compromised, thus proteins are hyperubiquitinated with K63-linked chains. With sufficient linear ubiquitination, the infrastructure of the signalling complex is sustained, caspase recruitment to TNFR1 is prohibited, and pro-survival signalling is enhanced.

**c, In A20 OUT(C103A) cells** with deficient linear ubiquitination, removal of K63 ubiquitin chains is still compromised; however, decreased linear chains favours enhanced association of hyperubiquitinated RIPK1 with FADD and caspase 8, the proximal components of the pro-death complex. Enhanced caspase 8 recruitment and activation in turn activates downstream effector caspases (such as caspase 3 and 7), culminating in cell death.



**Extended Data Figure 2 | Engineering and genotyping of A20 OTU mutant and A20 ZnF4 mutant knock-in mice.** **a.** Schematic diagrams of the A20 protein indicating the locations of the knock-in point mutations for each engineered mouse strain. The gene and protein names are also indicated, with abbreviated protein names indicated in parentheses. **b.** *Tnfaip3*<sup>Otu/Otu</sup> knock-in allele encoding A20 OTU(C103A) and representative genotyping data (right panel). **c.** *Tnfaip3*<sup>z4Cys/z4Cys</sup> knock-in

allele encoding A20 ZnF4(C609A,C612A) and representative genotyping data (right panel). **d.** *Tnfaip3*<sup>z4Ub/z4Ub</sup> knock-in allele encoding A20 ZnF4(Y599A,F600A) and representative genotyping data (right panel). **b–d.** Correctly targeted ES cell clones were identified by long-range PCR followed by sequencing (data not shown). LoxP sites are illustrated as yellow arrows, frt sites as red arrows. Modified exons 3 and 7 are indicated in blue. For gel source data, see Supplementary Fig. 6.



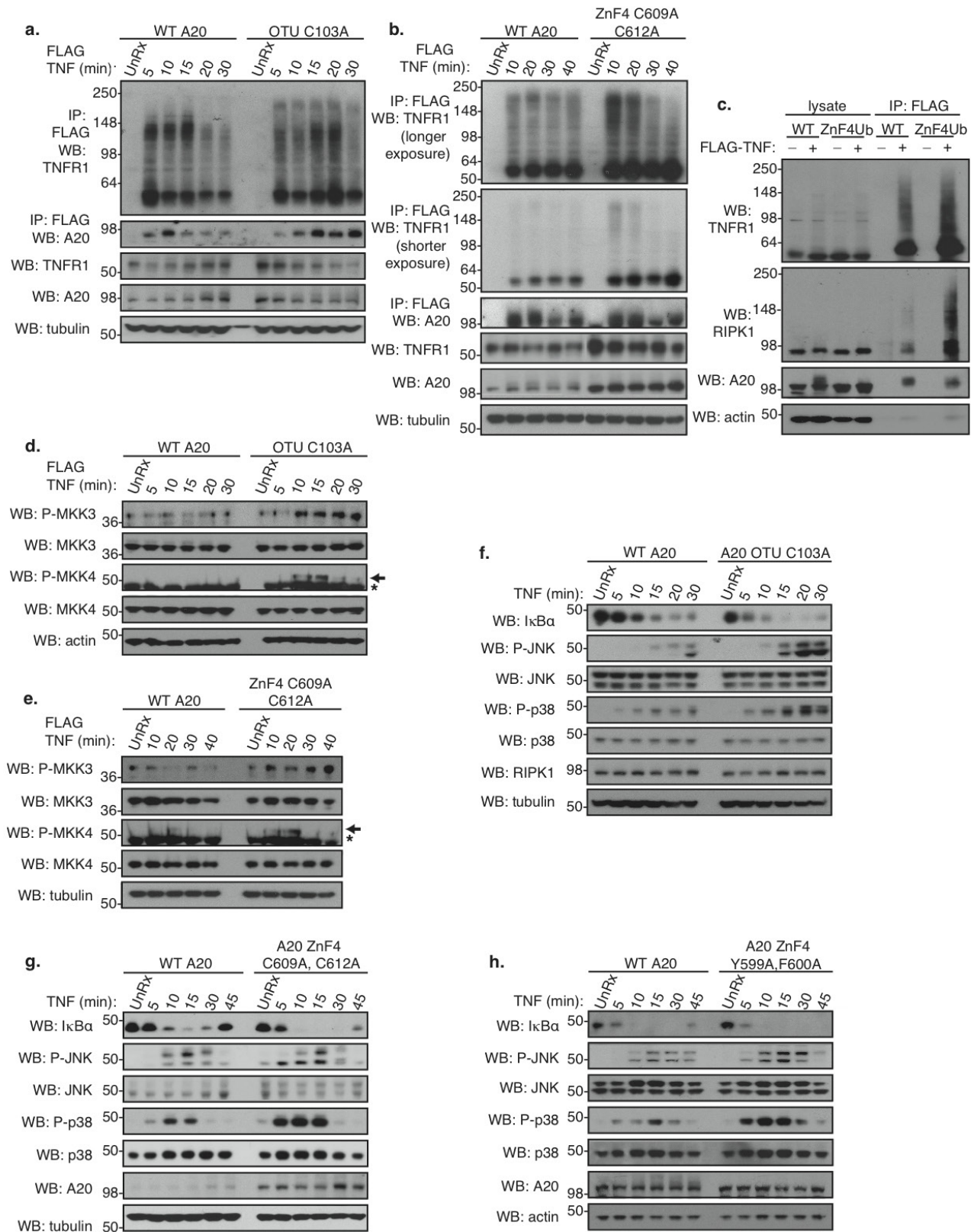
Extended Data Figure 3 | See next page for caption.



**Extended Data Figure 3 | Analysis of TNF-challenged A20****wild-type, A20 OTU(C103A), A20 ZnF4(C609A,C612A) and A20**

**ZnF4(Y599A,F600A) mice.** **a**, Body temperatures of mice in response to 300 µg TNF per kg body weight treatment. Error bars are indicated for each data point and represent the mean  $\pm$  standard deviation of 3 or 4 mice per genotype. **b**, A heat map representing profiles of serum cytokines in 12 different genotype/TNF treatment groups. Mice ( $n = 3$  or 4 per group) were treated for the indicated time with 300 µg TNF per kg body weight; mean values per group are represented in the heat map. Each row represents one cytokine, whose values were standardized to z-scores with a mean of zero and a standard deviation of 1, and colour-coded according to the colour key. Variances of selected serum cytokines from A20 wild-type, A20 OTU(C103A) (OTU), or A20 ZnF4(C609A,C612A) (ZnF4 Cys) mice in response to TNF stimulation were evaluated using the Student's *t*-test: IL6 WT versus ZnF4 Cys 2 h  $P = 0.009$ , 4 h  $P = 0.030$ ; IL6 WT versus OTU 2 h  $P = 0.023$ , 4 h  $P = 0.035$ ; Cxcl1 WT versus ZnF4 Cys 2 h  $P = 0.011$ , 4 h  $P = 0.017$ ; Cxcl1 WT versus OTU 2 h  $P = 0.047$ , 4 h  $P = 0.043$ ; Csf3 WT versus ZnF4 Cys 4 h  $P = 0.017$ , Csf3 WT versus OTU 4 h  $P = 0.042$ ; Ccl11 WT versus ZnF4 Cys 4 h  $P = 0.0003$ , Ccl11 WT versus OTU 4 h  $P = 0.036$ . **c**, Body temperatures of ZnF4 mutant mice in response to 300 µg TNF per kg body weight treatment. Error bars are indicated for each data point and represent the mean  $\pm$  standard deviation of 4 mice per genotype. **d**, A heat map representing profiles of serum cytokines as in Extended Data Figure 3b, but the indicated mice ( $n = 4$  per group) were treated for four hours with 300 µg TNF per kg body weight or PBS vehicle control. Variances of

selected serum cytokines from A20 ZnF4(C609A,C612A) (ZnF4 Cys), A20 ZnF4(Y599A,F600A) (ZnF4 Ub) or the respective wild-type control mice in response to TNF stimulation were evaluated using the Student's *t*-test: IL6 WT versus ZnF4 Cys  $P = 0.000057$ ; IL6 WT versus ZnF4 Ub  $P = 0.014$ ; Cxcl1 WT versus ZnF4 Cys  $P = 0.016$ ; Cxcl1 WT versus ZnF4 Ub  $P = 0.012$ ; Csf3 WT versus ZnF4 Cys  $P = 0.024$ , Csf3 WT versus ZnF4 Ub  $P = 0.0061$ ; Ccl11 WT versus ZnF4 Cys  $P = 0.005$ , Ccl11 WT versus ZnF4 Ub  $P = 0.001$ . **e**, Analysis of myelin oligodendrocyte glycoprotein-induced experimental autoimmune encephalomyelitis (MOG-EAE) studies in A20 WT, A20 OTU(C103A), and A20 ZnF4(C609A,C612A) mice. Top panel, MOG-EAE disease scores over time (mean  $\pm$  s.e.m.) for A20 WT ( $n = 15$ ) and A20 OTU(C103A) ( $n = 14$ ). A20 OTU(C103A) average daily clinical scores (ADCS)  $P = 0.012$ , Dunnett's test versus A20 WT. Bottom panel, EAE disease scores over time (mean  $\pm$  s.e.m.) for A20 WT ( $n = 13$ ) and A20 ZnF4(C609A,C612A) ( $n = 12$ ). A20 ZnF4(C609A,C612A) ADCS  $P = 0.046$ , Dunnett's test versus A20 WT. **f**, Lower power (upper panel A) and higher power (lower panel B) microscopic images of a representative lumbar spinal cord section derived from a A20 ZnF4(C609A,C612A) mouse with a grade 3 EAE clinical score at study termination (day 30). The section is stained with haematoxylin, eosin and Luxol fast blue. **A**, Foci of myelinopathy and gliosis (arrows). Gr, grey matter; Wh, white matter; scale bar, 200 µm. **B**, Focally severe myelinopathy and gliosis (double arrows) extending from the meninges close to the grey matter (delineated by white dashed line). Scale bar, 50 µm. Data represent at least two biological replicates.

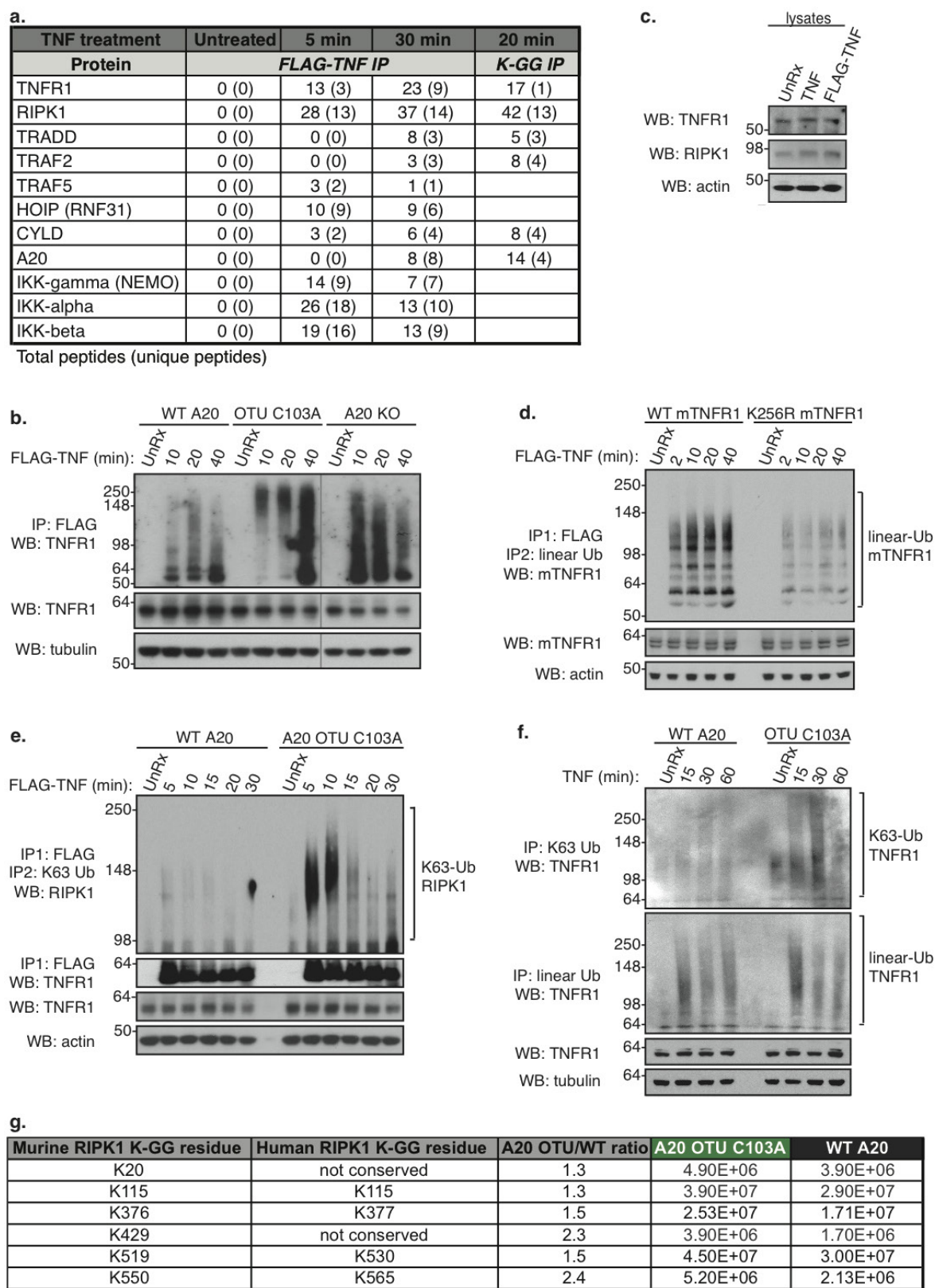


Extended Data Figure 4 | See next page for caption.

**Extended Data Figure 4 | A20 proteins from wild-type, A20 OTU(C103A) cells, A20 ZnF4(C609A,C612A) cells, and A20 ZnF4(Y599A,F600A) cells are efficiently recruited to TNFR1 and regulate downstream signalling.** **a**, Immunoblot analysis Flag–TNF-engaged immunocomplexes and the corresponding whole-cell lysates in wild-type and A20 OTU(C103A) MEFs. **b**, Immunoblot analysis Flag–TNF-engaged immunocomplexes and the corresponding whole-cell lysates in wild-type and A20 ZnF4(C609A,C612A) MEFs. **c**, Immunoblot analysis Flag–TNF-engaged immunocomplexes and the corresponding whole-cell lysates in wild-type and A20 ZnF4(Y599A,F600A) MEFs. **d**, Immunoblot analysis Flag–TNF-treated whole-cell lysates in wild-type and in A20 OTU(C103A) MEFs. Asterisk, background band; arrow, phospho-MKK4. UnRx, untreated. **e**, Immunoblot analysis Flag–TNF-treated whole-cell lysates in wild-type and in A20 ZnF4(C609A,C612A) MEFs. Asterisk, background band; arrow, phospho-MKK4. UnRx, untreated. **f**, Immunoblot analysis of whole-cell lysates from TNF-treated

wild-type and A20 OTU(C103A) MEFs. Immunoblot analysis of whole-cell lysates from TNF-treated wild-type, A20 OTU(C103A), and A20 null MEFs following TNF pre-treatment to induce A20 expression, as well as TNF-treated A20 wild-type and A20 OTU(C103A) primary BMDMs, and TNF-treated A20 wild-type and A20 OTU(C103A) immortalized BMDMs all showed similar trends (not shown). **g**, Immunoblot analysis of whole-cell lysates from TNF-treated wild-type and A20 ZnF4 C609,612A E1A transformed MEFs. UnRx, untreated. Immunoblot analysis of whole-cell lysates from TNF-treated wild-type and A20 ZnF4 C609,612A MEFs following TNF pre-treatment to induce A20 expression, and analysis of whole-cell lysates from TNF-treated A20 wild-type and A20 ZnF4 C609,612A primary BMDMs all showed similar trends (not shown). **h**, Immunoblot analysis of whole-cell lysates from TNF-treated wild-type and A20 ZnF4(Y599A,F600A) primary MEFs. For gel source data, see Supplementary Figs 6, 7. Data represent two to four biological replicates.

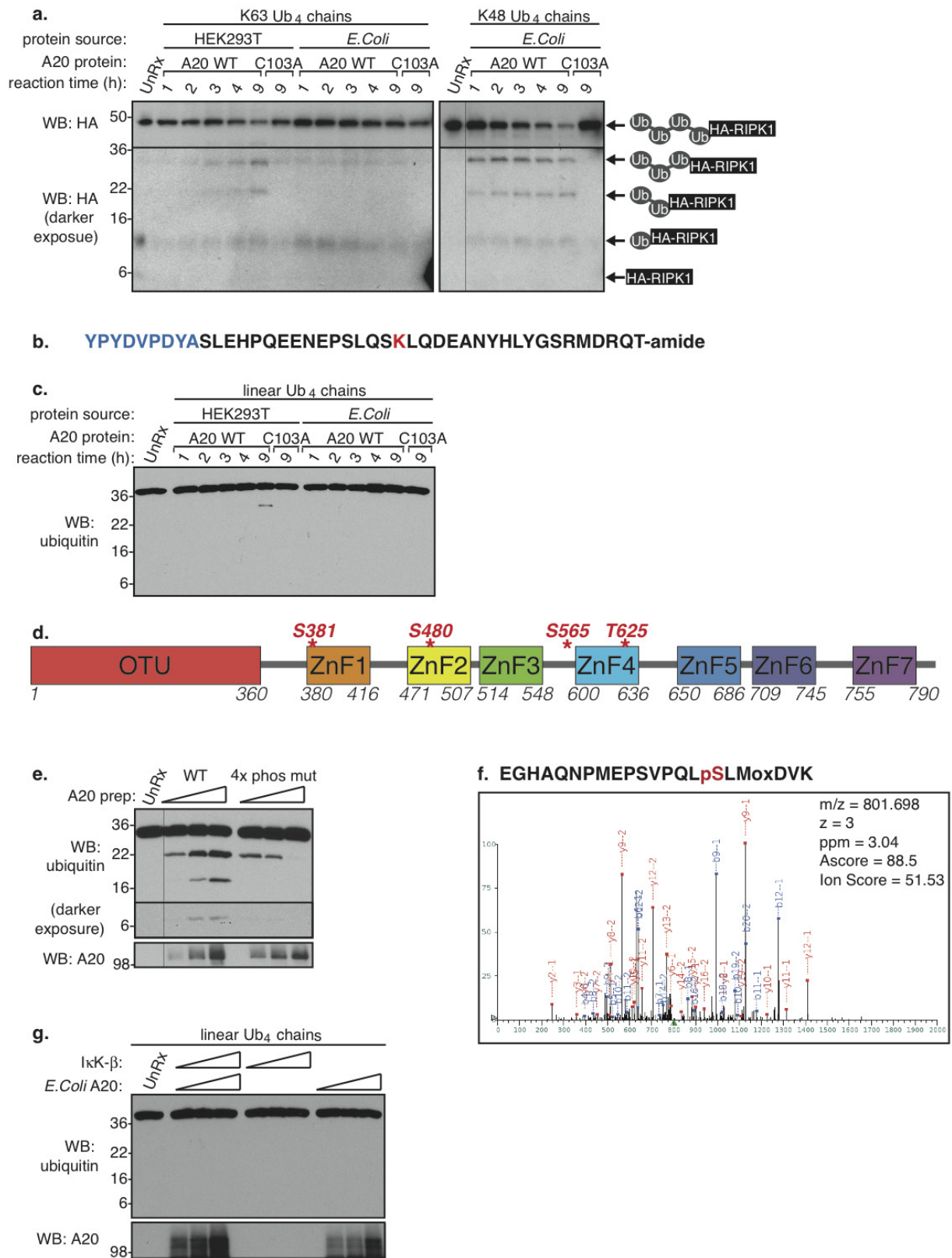




Extended Data Figure 5 | See next page for caption.

**Extended Data Figure 5 | Additional analysis of ubiquitination status analysis of TNFR1 and associated proteins.** **a**, A summary table of selected proteins identified in anti-Flag immunocomplexes from untreated or in Flag–TNF-treated wild-type MEFs by LC-MS/MS analysis (left columns) and a summary of the ubiquitination sites identified on the indicated proteins from TNF-treated A20 wild-type MEFs with the PTMscan approach using anti-K- $\epsilon$ -GG antibodies and LC-MS/MS (right column). **b**, Analysis of TNF-engaged TNFR1 in untreated or Flag–TNF-treated E1A-transformed A20 wild-type or A20 OTU(C103A) MEFs, or in A20 null primary MEFs. Anti-Flag immunocomplexes were purified using Flag peptide elution and elutions were blotted for TNFR1. Immunoblots of the corresponding whole-cell lysates are indicated below. **c**, Lysates corresponding with Fig. 1c. **d**, Murine TNFR1(K256R) attenuates TNFR1 ubiquitination and downstream signalling. Murine wild-type or TNFR1(K256R) was transfected in human 293T cells and cells were treated with Flag–TNF as indicated. Equal inputs of lysates were immunoprecipitated with anti-Flag, dissociated and

re-immunoprecipitated with anti-linear ubiquitin antibody, and blotted for murine TNFR1, or lysates were blotted with the indicated antibodies. **e**, Analysis of TNFR1-associated RIPK1 K63 ubiquitination (Ub) status and TNFR1 immunoprecipitates in Flag–TNF-treated wild-type and A20 OTU(C103A) MEFs. **f**, Comparison of activated TNFR1 ubiquitination status in E1A transformed A20 wild-type and OTU(C103A) MEFs. Treated cells were lysed in buffer containing 6 M urea and immunoprecipitated with the indicated antibodies under denaturing conditions. **g**, Summary table of RIPK1 ubiquitination sites identified from TNF-treated A20 wild-type and OTU(C103A) MEFs with the PTMscan approach using anti-K- $\epsilon$ -GG antibodies and LC-MS/MS. Peptides were quantified with area under curve (AUC) and summarized to site level. The equivalent human RIPK1 residues are also indicated. The average ratio of endogenous RIPK1 ubiquitination sites in A20 OTU(C103A): wild-type A20 is 1.7. Additional TNFR1 mass spectrometry data are shown in Supplementary Information a–c. For gel source data, see Supplementary Figs 7, 8. Data represent two to four biological replicates.

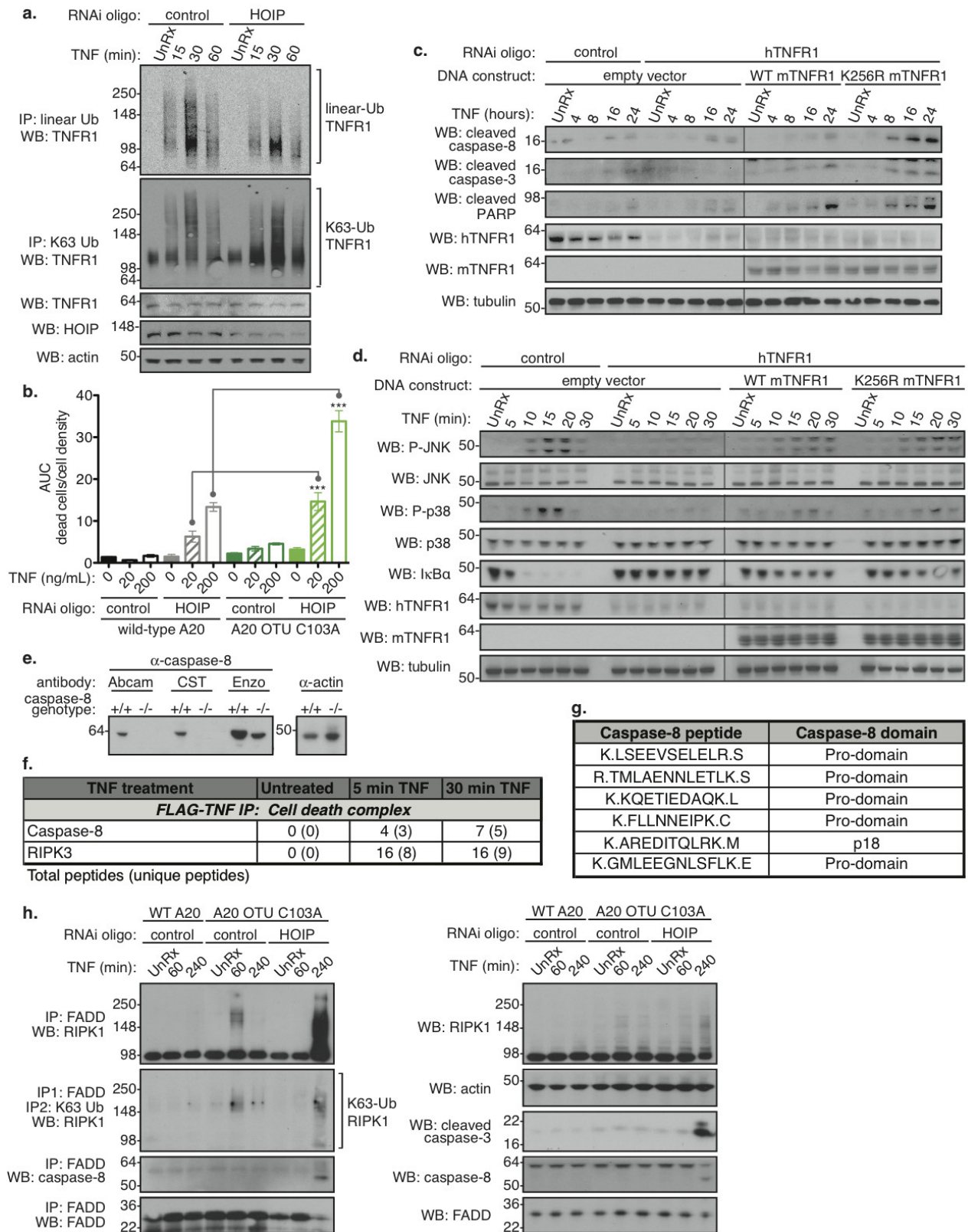


Extended Data Figure 6 | See next page for caption.



**Extended Data Figure 6 | *In vitro* deubiquitination assays.** Additional data corresponding to Fig. 2. Normalized A20 WT or OTU(C103A) inputs purified from *E. coli* or mammalian HEK 293T cells are shown in Fig. 2a. **a**, A20-mediated cleavage efficacy of K63- or K48-linked tetraubiquitin conjugated to a HA-tagged RIPK1 peptide. **b**, Sequence of the HA epitope-tagged human RIPK1 peptide. The HA epitope tag is shown in blue, the human RIPK1 residues in black, and K377 is highlighted in red. **c**, Cleavage time course of linear tetraubiquitin by purified A20 WT or OTU(C103A) from *E. coli* or from mammalian HEK 293T cells. Input protein levels are shown in Fig. 2a. **d**, A schematic of the human A20 protein indicating where the phosphorylation sites are localized. Mass spectrometry PhosphoSite analysis (<http://www.phosphosite.org/>) of

A20 derived from mammalian expression systems is shown in Supplementary Information d. **e**, Comparison of the cleavage efficacy of K63-linked tetraubiquitin with increasing doses of human wild-type A20 or phospho-site mutant A20 (4× phos mut). 4× phos mut: S381A, S480A, S565A, and T625A. Wild-type or phos mut A20 proteins were expressed in and purified from mammalian HEK 293T cells. **f**, Tandem mass spectrum for the S381-containing peptide from human A20 expressed in *E. coli* and phosphorylated with recombinant I $\kappa$ K $\beta$ . **g**, Cleavage efficacy of linear tetraubiquitin chains by increasing doses of *E. coli*-derived wild-type A20, I $\kappa$ K $\beta$  alone, or I $\kappa$ K $\beta$ -phosphorylated A20. For gel source data, see Supplementary Figs 8, 9. Data represent two to five biological replicates.

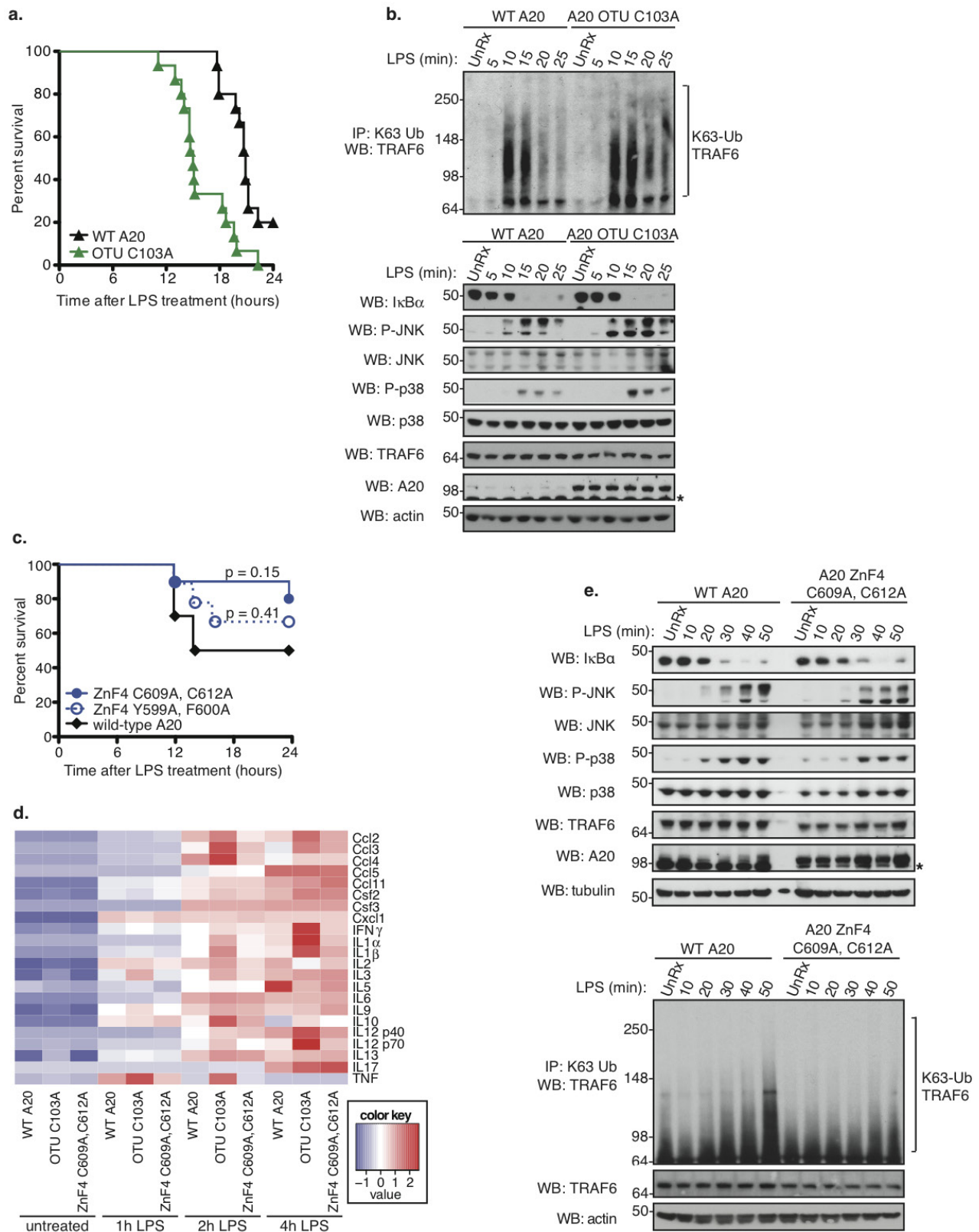


Extended Data Figure 7 | See next page for caption.

**Extended Data Figure 7 | Effects of linear ubiquitination in modulating TNFR1 signalling and cell viability.** **a**, HOIP RNAi decreases linear ubiquitination of TNFR1. Wild-type MEFs were transfected with control or HOIP RNAi oligonucleotides and treated for the indicated times with TNF. Treated cells were lysed in buffer containing 6 M urea and immunoprecipitated with the indicated antibodies under denaturing conditions. Immunoprecipitates and whole-cell lysates were blotted as indicated. **b**, Area under the curve (AUC) data corresponding to cell death data in Fig. 3a. All error bars are s.e.m. for technical triplicates. \*\*\* $P < 0.001$  determined by *t*-test. **c**, Murine TNFR1(K256R) enhances caspase activation. Human HEK 293T cells were treated with control or with human TNFR1 RNAi oligonucleotides and transfected with murine wild-type or with TNFR1(K256R) as indicated. Cells were treated with TNF as indicated and equal inputs of lysates were immunoblotted with the indicated antibodies. **d**, Murine TNFR1(K256R) does not modulate MAPK or NF- $\kappa$ B signalling. Human HEK 293T cells were treated with control or with human TNFR1 RNAi oligonucleotides and transfected with murine wild-type or with TNFR1(K256R) as indicated. Cells were treated with

TNF as indicated and equal inputs of lysates were immunoblotted with the indicated antibodies. **e**, Evaluation of the specificity of anti-caspase 8 antibodies. E1A-transformed MEFs of the indicated genotype were lysed in buffer containing 6 M urea, quantified, and immunoblotted with the indicated antibodies as detailed in the Methods. **f**, A summary table of cell death proteins identified in anti-Flag immunocomplexes from untreated or in Flag-TNF-treated wild-type MEFs by LC-MS/MS analysis (also see Supplementary Fig. 6a). **g**, Sequences of the caspase 8 peptides in anti-Flag immunocomplexes from untreated or in Flag-TNF-treated wild-type MEFs by LC-MS/MS analysis. **h**, Left panels, analysis of FADD immunoprecipitates and FADD-associated RIPK1 K63 ubiquitination (Ub) status in TNF-treated wild-type and A20 OTU(C103A) MEFs transfected with control- or HOIP RNAi oligonucleotides. Right panels, immunoblot analysis of whole-cell lysates from TNF-treated wild-type and A20 OTU(C103A) MEFs transfected with control- or HOIP RNAi oligonucleotides. HOIP knockdown was validated by RT-PCR analysis (data not shown). For gel source data, see Supplementary Figs 9, 10. Data represent two to three biological replicates.





Extended Data Figure 8 | See next page for caption.

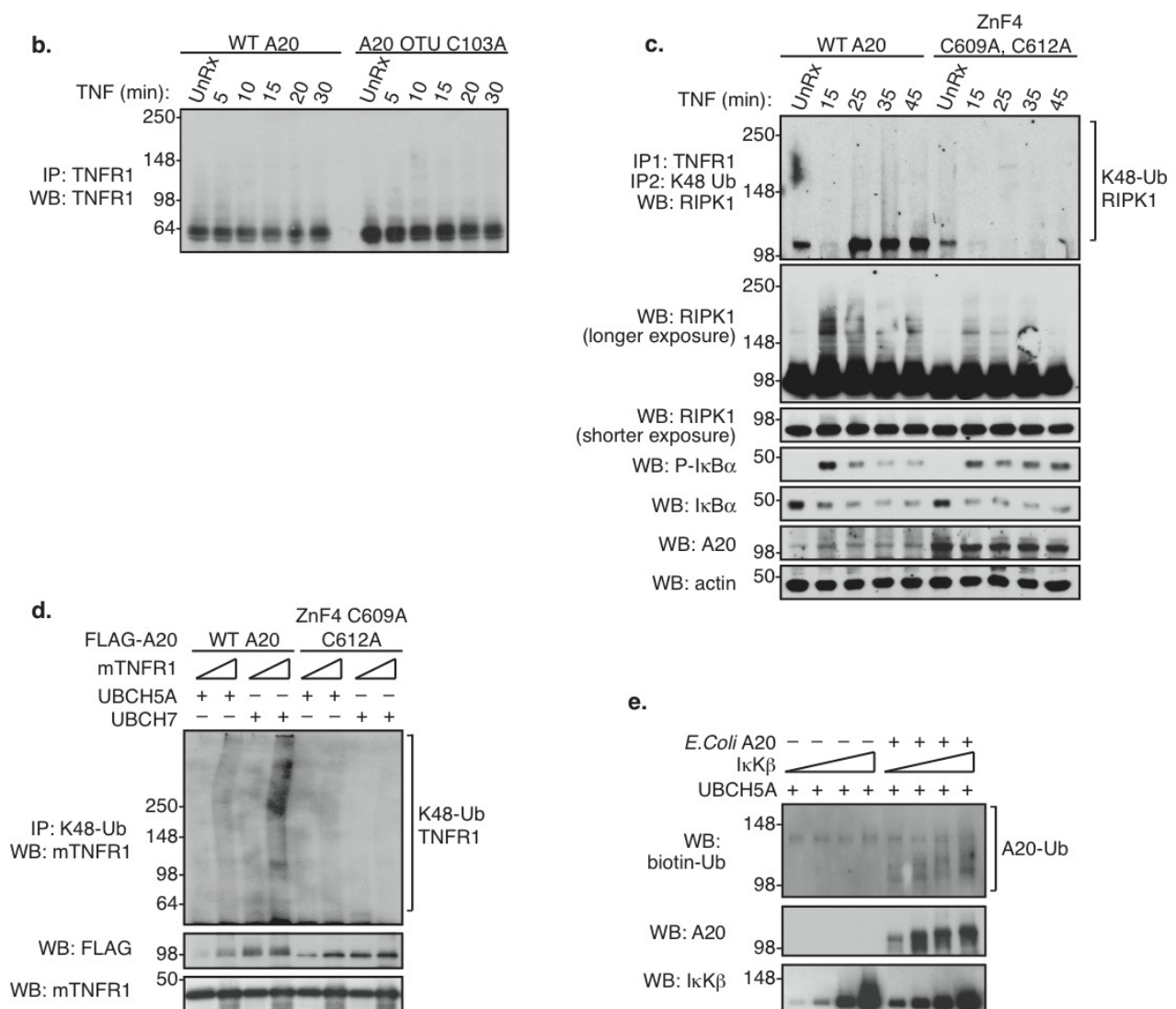
**Extended Data Figure 8 | A20 OTU domain, but not the ZnF4 motif, downmodulates LPS signalling.** **a**, Kaplan–Meier survival curves of A20 WT ( $n = 15$ ) and A20 OTU(C103A) ( $n = 15$ ) mice in response to 20 mg LPS per kg body weight. Log rank  $P = 0.0002$ , Wilcoxon  $P < 0.0001$ . **b**, Upper panel, analysis of TRAF6 K63 ubiquitination (Ub) status in LPS-treated WT and A20 OTU(C103A) primary BMDMs. Lower panels, immunoblot analysis of whole-cell lysates from LPS-treated wild-type and A20 OTU(C103A) primary BMDMs. Asterisk, background band. Similar trends were seen in wild-type and A20 OTU(C103A) MEFs in response to acute LPS treatment and following LPS pre-treatment to induce A20 expression (not shown). **c**, Kaplan–Meier survival curves of A20 wild-type ( $n = 10$ ), A20 ZnF4(C609A,C612A) ( $n = 10$ ), and A20 ZnF4(Y599A,F600A) ( $n = 9$ ) mice in response to 20 mg LPS per kg body weight. Log rank  $P = 0.1531$ , Wilcoxon  $P = 0.1398$  for A20 ZnF4(C609A,C612A) versus WT; Log rank  $P = 0.4103$ , Wilcoxon  $P = 0.3373$  for A20 ZnF4(Y599A,F600A) versus WT. **d**, A heat map representing profiles of serum cytokines in 12 different genotype/LPS treatment groups. Mice ( $n = 3$  or 4 per group) were treated for the indicated time with 40 mg LPS per kg body weight LPS as indicated; mean values per group are represented in the heat map. Each row

represents one cytokine, whose values were standardized to z-scores with a mean of zero and a standard deviation of 1, and colour-coded according to the colour key. Variances of selected serum cytokines from A20 WT, A20 OTU(C103A) (OTU), or A20 ZnF4(C609A,C612A) (ZnF4 Cys) mice in response to LPS stimulation were evaluated using the Student's *t*-test: TNF WT versus OTU 2 h  $P = 0.044$ ; IFN $\gamma$  WT versus OTU 4 h  $P = 0.033$ ; Ccl4 WT versus OTU 4 h  $P = 0.040$ . Profiles of selected serum cytokines from A20 WT ( $n = 5$ ), ZnF4 Cys ( $n = 4$ ), or OTU(C103A) ( $n = 5$ ) mice in response to PBS or low dose (5 mg LPS per kg body weight) LPS and collected at 2 h or 6 h post-stimulation showed similarly significant variances between A20 WT and A20 OTU(C103A) (OTU) but not between A20 WT and A20 ZnF4(C609A,C612A) (ZnF4 Cys) (not shown). **e**, Upper panel, immunoblot analysis of whole-cell lysates from LPS-treated wild-type A20 and A20 ZnF4 C609,612A MEFs. Lower panel, analysis of TRAF6 K63 ubiquitination (Ub) status in the corresponding MEFs. Similar trends were seen in lysates from LPS-treated wild-type and A20 ZnF4 C609,612A MEFs following LPS pre-treatment to induce A20 expression (not shown). For gel source data, see Supplementary Fig. 10. Data represent two biological replicates.

a.

A20 ZnF	A20 residues		K <sub>D</sub> (μM)			
	human	murine	mono-Ub	tri-Ub K63	tri-Ub linear	tri-Ub K48
ZnF4	WT	WT	230 ± 20	1.7 ± 0.1	> 50	N/D
ZnF4	C624A, C627A	C609A, C612A	N/D	N/D	N/D	N/D
ZnF4	site I I629R		90 ± 30			
ZnF4	site II Y614A, F615A	Y599A, F600A	N/D	N/D	N/D	N/D
ZnF4	site III K606E		110 ± 30			
ZnF4	site I+III I629R, K606E		50 ± 13			
ZnF7	WT		110 ± 40	N/D	0.004 ± 0.002*	N/D
ZnF7	F770A, G771A		N/D	N/D	N/D	N/D
ZnF1	WT		N/D	N/D	N/D	N/D

N/D: no detected binding

\*Value is for K<sub>D</sub> strong; K<sub>D</sub> weak = 0.5 ± 0.1 μM

Extended Data Figure 9 | See next page for caption.



**Extended Data Figure 9 | Characterization of A20 ZnF mutants and their cellular effects.** **a**, Summary of binding data of wild-type human A20 ZnF motifs and ZnF mutants to mono-ubiquitin, as measured by NMR, and to tri-ubiquitin chains, as measured by biolayer interferometry. Data are shown in Fig. 5b and Supplementary Information h, g. **b**, Analysis of TNF-engaged TNFR1 in untreated or TNF-treated E1A transformed A20 wild-type or A20 OTU(C103A) MEFs. Anti-TNF immunocomplexes were captured using anti-TNFR1 antibody-coupled beads and elutions were blotted for TNFR1. **c**, Analysis of TNFR1-associated RIPK1 K48 ubiquitination (Ub) status in TNF-treated wild-type and A20 ZnF4(C609A,C612A) MEFs. Immunoblot analysis

of the corresponding whole-cell lysates are shown in the lower panels. **d**, Flag-wild-type A20 ubiquitinates recombinant murine TNFR1 with K48 chains. Flag-wild-type A20 or Flag-A20 ZnF4(C609A,C612A) proteins purified from HEK 293T lysates were added to *in vitro* reactions with recombinant murine TNFR1 and ubiquitin system enzymes. Reactions were immunoprecipitated in 4 M urea using an anti-K48 ubiquitin antibody, and immunoblotted, or reaction inputs were blotted as indicated. **e**, I $\kappa$ K $\beta$ -phosphorylated A20, but not I $\kappa$ K $\beta$  alone, promotes *in vitro* ubiquitination. For gel source data, see Supplementary Fig. 11. Data represent two to three biological replicates.

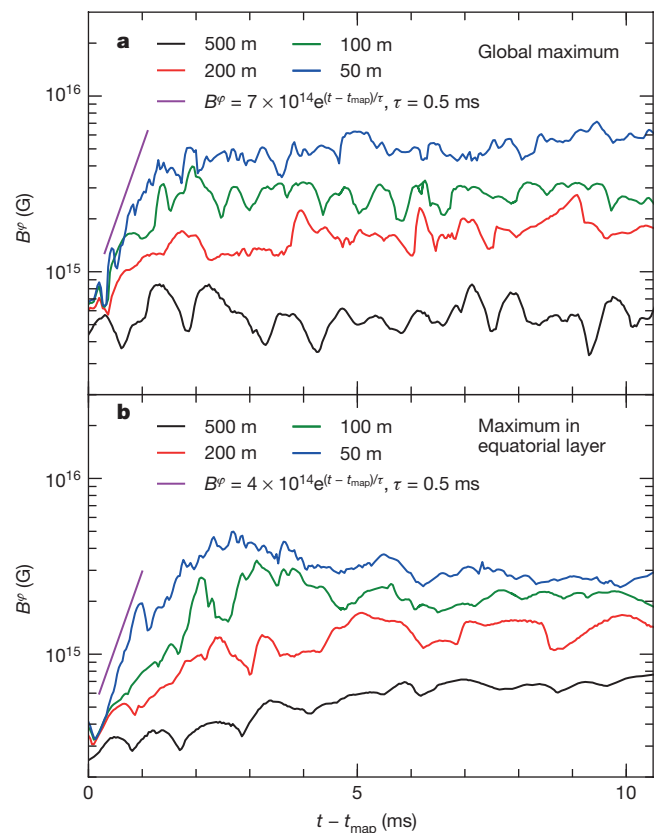
# A large-scale dynamo and magnetoturbulence in rapidly rotating core-collapse supernovae

Philipp Mösta<sup>1,2</sup>, Christian D. Ott<sup>1</sup>, David Radice<sup>1</sup>, Luke F. Roberts<sup>1</sup>, Erik Schnetter<sup>3,4,5</sup> & Roland Haas<sup>6</sup>

Magnetohydrodynamic turbulence is important in many high-energy astrophysical systems, where instabilities can amplify the local magnetic field over very short timescales<sup>1,2</sup>. Specifically, the magnetorotational instability and dynamo action<sup>3–6</sup> have been suggested as a mechanism for the growth of magnetar-strength magnetic fields (of  $10^{15}$  gauss and above) and for powering the explosion<sup>7–10</sup> of a rotating massive star<sup>11,12</sup>. Such stars are candidate progenitors of type Ic-bl hypernovae<sup>13,14</sup>, which make up all supernovae that are connected to long  $\gamma$ -ray bursts<sup>15,16</sup>. The magnetorotational instability has been studied with local high-resolution shearing-box simulations in three dimensions<sup>17–19</sup>, and with global two-dimensional simulations<sup>20</sup>, but it is not known whether turbulence driven by this instability can result in the creation of a large-scale, ordered and dynamically relevant field. Here we report results from global, three-dimensional, general-relativistic magnetohydrodynamic turbulence simulations. We show that hydromagnetic turbulence in rapidly rotating protoneutron stars produces an inverse cascade of energy. We find a large-scale, ordered toroidal field that is consistent with the formation of bipolar magnetorotationally driven outflows. Our results demonstrate that rapidly rotating massive stars are plausible progenitors for both type Ic-bl supernovae<sup>13,21,22</sup> and long  $\gamma$ -ray bursts, and provide a viable mechanism for the formation of magnetars<sup>23,24</sup>. Moreover, our findings suggest that rapidly rotating massive stars might lie behind potentially magnetar-powered superluminous supernovae<sup>25,26</sup>.

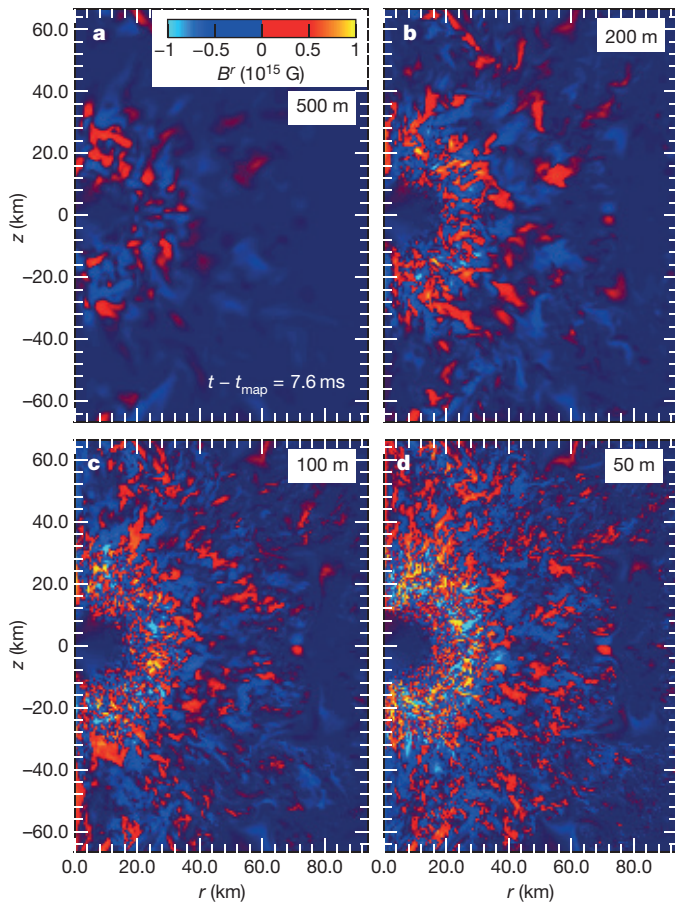
We study magnetohydrodynamic (MHD) turbulence in the shear layer around a rapidly rotating protoneutron star by using high-resolution, global, three-dimensional, general-relativistic (GR) MHD simulations (the resolution is about ten times higher than that of previous simulations). We take our initial conditions from a full three-dimensional GRMHD adaptive-mesh-refinement simulation<sup>9</sup> of stellar collapse in a rapidly spinning progenitor star. The initial spin period of the iron core,  $P_0$ , is 2.25 s before collapse; the spin period of the protoneutron star after core bounce (when the collapsing core rebounds, launching the initial shock wave),  $P_{\text{PNS}}$ , is 1.18 ms; and the initial maximum magnetic field is  $10^{10}$  G. We map to a high-resolution domain at time  $t_{\text{map}} = 20$  ms after core bounce. At this time, flux compression and linear winding<sup>27</sup> has built up a maximum toroidal field of about  $7 \times 10^{14}$  G close to the rotation axis of the protoneutron star, and about  $3 \times 10^{14}$  G in the equatorial region. The maximum poloidal magnetic field is about  $7 \times 10^{14}$  G at  $t_{\text{map}} = 20$  ms after core bounce. We carry out simulations at four resolutions,  $dx = [500 \text{ m}, 200 \text{ m}, 100 \text{ m}, 50 \text{ m}]$ ; adopt a domain size of 66.5 km in the  $x$  and  $y$  directions and 133 km in the  $z$  direction (rotation axis); and use a  $90^\circ$  rotational symmetry in the  $x$ – $y$  plane (with no symmetry in the  $z$  plane). This allows us to study the shear layer surrounding the core of the protoneutron star with unprecedented resolution, using fully self-consistent global three-dimensional simulations of MHD turbulence in stellar collapse.

The two lowest-resolution simulations show no or only minor amplification of the toroidal magnetic field, consistent with not resolving the fastest-growing mode (FGM) of the magnetorotational instability (MRI). The toroidal field in the two highest-resolution simulations exhibits exponential growth soon after the start of the simulations (Fig. 1). The poloidal magnetic field evolution follows the toroidal one closely (Extended Data Fig. 1). The initial transition to exponential growth in the global maximum toroidal field (Fig. 1a), and in the maximum toroidal field in a box with height 7.5 km above and below the equatorial plane (Fig. 1b), is nearly identical between the 100-m and the 50-m simulations. This indicates that we can resolve the FGM of the



**Figure 1 | Evolution of the maximum toroidal magnetic field.** Both panels show the maximum toroidal magnetic field ( $B^\phi$ ) as a function of time for the four resolutions 500 m, 200 m, 100 m and 50 m. **a**, The global maximum field; **b**, the maximum field in a thin layer above and below the equatorial plane ( $-7.5 \text{ km} \leq z \leq 7.5 \text{ km}$ ). The magenta line indicates exponential growth with an exponential-folding time of  $\tau = 0.5$  ms.

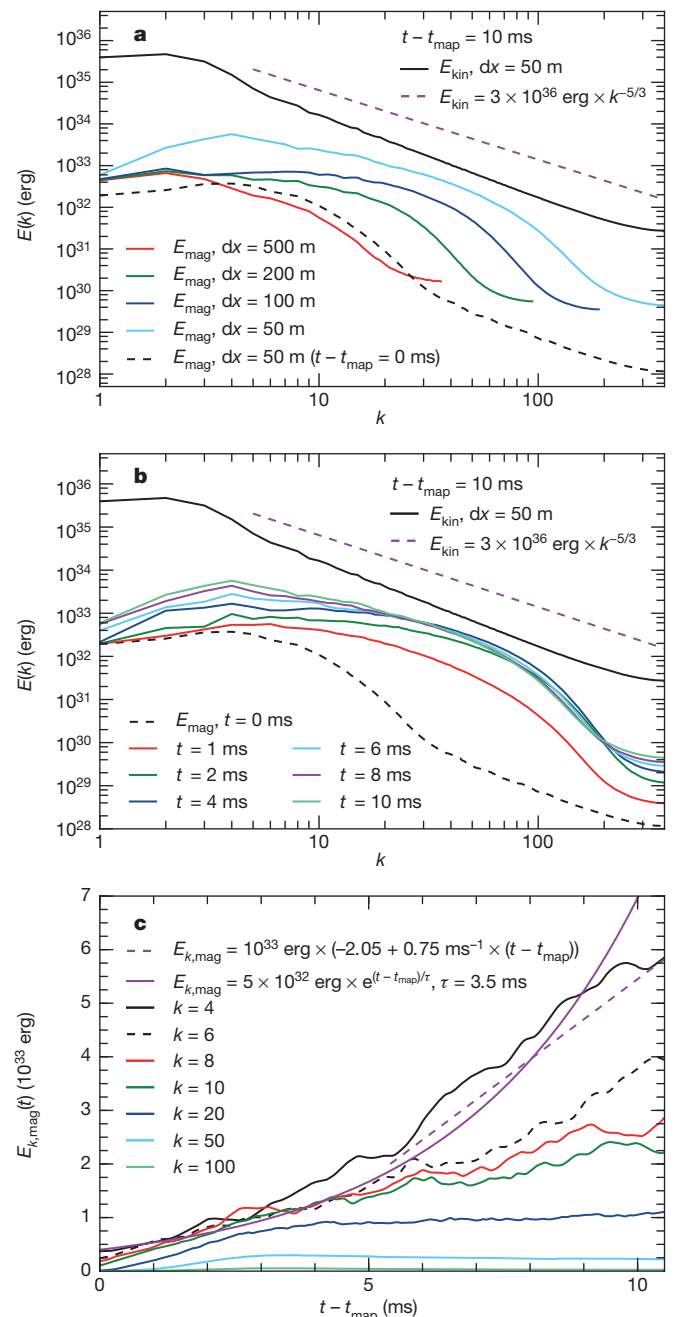
<sup>1</sup>TAPIR, Walter Burke Institute for Theoretical Physics, Mailcode 350-17, California Institute of Technology, Pasadena, California 91125, USA. <sup>2</sup>Department of Astronomy, 501 Campbell Hall #3411, University of California at Berkeley, Berkeley, California 94720, USA. <sup>3</sup>Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada. <sup>4</sup>Department of Physics, University of Guelph, Guelph, Ontario N1G 2W1, Canada. <sup>5</sup>Center for Computation & Technology, Louisiana State University, Baton Rouge, Louisiana, 70803, USA. <sup>6</sup>Max Planck Institute for Gravitational Physics, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany.



**Figure 2 | Radial magnetic field strength.** a–d, Visualization of the radial component of the magnetic field ( $B^r$ ) in two-dimensional  $r$ - $z$  slices at azimuth =  $45^\circ$ , for the four resolutions 500 m (a), 200 m (b), 100 m (c) and 50 m (d), at  $t - t_{\text{map}} = 7.6$  ms. The colour map ranges from positive  $10^{15}$  G (yellow) to negative  $10^{15}$  G (light blue).

MRI with the 100-m simulation, and is consistent with our background flow stability analysis of the initial adaptive-mesh-refinement simulation (Extended Data Fig. 2). The observed growth time of  $\tau \approx 0.5$  ms agrees well with the analytically predicted growth time of the FGM from linear analysis. The field evolution quickly becomes nonlinear, and this rapid growth reaches a fully turbulent saturated state within 3 ms. The turbulent saturated toroidal field strength agrees to within a factor of two between the two highest-resolution simulations (100 m and 50 m). Once nonlinear field strength is reached, secondary modes and couplings between individual modes become important for the observed growth time of the MRI. The final turbulent saturation field is not converged and differs between resolutions, because secondary instabilities, resistivity, and finite resolution effects become important<sup>28,29</sup>. However, these differences decrease with increasing resolution and we expect our results to hold when even higher-resolution simulations become computationally accessible. This expectation is supported by the fact that the local features of our global three-dimensional simulations are consistent with previous higher-resolution ( $dx \approx 10$  m) local simulations of the MRI<sup>17</sup>.

The resolution dependence of the magnetic field in the turbulent state is striking (Fig. 2). Although the 500-m and 200-m simulations show none to only mild turbulence, the 100-m and 50-m simulations develop a fully turbulent shear layer around the protoneutron star. We observe radial filaments of magnetic field that oscillate from negative to positive values on a length scale of 1 km, consistent with the predicted wavelength of the FGM of the MRI (Extended Data Fig. 2). These structures resemble the formation of channel flows that are observed in shearing-box simulations<sup>17</sup>, but do not stay coherent because of

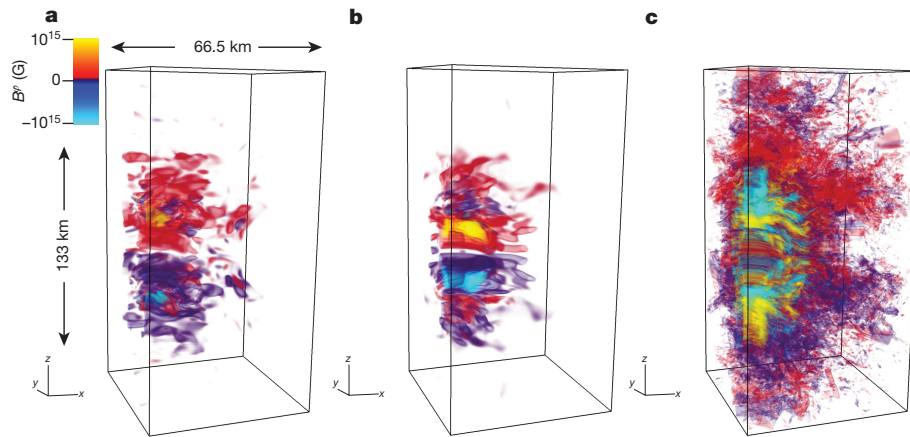


**Figure 3 | Turbulent kinetic and electromagnetic energy spectra.** a, b, Energy as a function of the dimensionless wavenumber  $k$  ( $E(k)$ ). Panel a compares the electromagnetic energy ( $E_{\text{mag}}$ ) across all four resolutions at  $t - t_{\text{map}} = 10$  ms. Panel b shows a time series of electromagnetic energy spectra for the 50-m simulation only. Both panels show the turbulent kinetic energy ( $E_{\text{kin}}$ ) as computed from the 50-m simulation (black solid line); a line indicating Kolmogorov scaling ( $k^{-5/3}$ ) (purple dashed line); and the initial electromagnetic energy spectrum (black dashed line). c, The electromagnetic energy at a given wavenumber ( $E_{k,\text{mag}}(t)$ ) versus time, and an exponential fit (purple solid line) and linear fit (purple dashed line).

the background flow. Similar non-coherent filaments have also been observed in two-dimensional global simulations<sup>20</sup>.

The turbulent kinetic and electromagnetic energy spectra calculated from our simulations are shown in Fig. 3. Initially, the turbulent kinetic energy, which is nearly constant in time, is several orders of magnitude larger across all scales than the electromagnetic energy. The electromagnetic energy is highly time and resolution dependent. Although the low-resolution calculation shows little evolution away





**Figure 4 | Three-dimensional volume renderings of the toroidal magnetic field,  $B^\varphi$ .** All panels show ray-casting volume renderings of  $B^\varphi$ . The rotation axis  $z$  is the vertical axis, and the volume renderings are generated using a varying-alpha colour map. Yellow indicates a positive field of strength  $10^{15}$  G; red denotes a weaker positive field; light blue

corresponds to a negative field of strength  $10^{15}$  G; darker blue indicates a weaker negative field. **a**, The initial conditions for our simulations; **b**, the 500-m simulation at time  $t - t_{\text{map}} = 10$  ms; **c**, the 50-m simulation at  $t - t_{\text{map}} = 10$  ms.

from the initial spectrum, the higher-resolution calculations saturate at larger and larger energies at large values of the dimensionless wavenumber  $k$  (Fig. 3a). The saturation value at large and intermediate  $k$  values is within a factor of three of equipartition with the turbulent kinetic energy in the 50-m calculation. After saturation is reached at large  $k$  values, we observe an inverse cascade of energy, which triggers the growth of large-scale electromagnetic energy peaking at  $k = 4$ , corresponding to a length scale of 5 km for our domain. This is well below the driving scale of the FGM of the MRI ( $k \approx 20$ ) and consistent with the structures evident in Figs 2d and 4c. The growth in the first 7 ms is fitted well by an exponential with exponential-folding time  $\tau = 3.5$  ms. We observe a transition away from clean exponential growth for  $t - t_{\text{map}} \geq 7$  ms; this transition might be caused by the magnetic field becoming dynamically relevant, and/or by (numerical) resistivity becoming important for the magnetic field evolution<sup>5</sup>. Here, the growth at  $k = 4$  is described better by a linear fit. In an inverse cascade, the energy is expected to reach approximately the same relative saturation value (with respect to the driving turbulent kinetic energy) at all  $k$  values with sufficiently long evolution times<sup>3,4</sup>. We find evidence for this in the range  $10 \leq k \leq 50$ , where the magnetic energy spectrum begins to evolve towards a similar power-law scaling as the turbulent kinetic energy. Assuming that this also holds at smaller  $k$  values, we can extrapolate the growth of magnetic energy on the basis of the linear fit (Fig. 3c). We expect to reach saturation electromagnetic energy at small  $k$  values within  $t - t_{\text{map}} \approx 60$  ms. The observed differences between the calculations for 100-m and 50-m resolution, in their saturation energies at large  $k$  values and in their inverse energy cascades, indicates that the turbulent state is not fully captured with the 100-m simulation and that the efficiency of the inverse cascade may still increase when going to even higher resolutions than 50 m.

Our results indicate that the electromagnetic energy will rival the turbulent kinetic energy and dominate the less efficient neutrino heating<sup>8,30</sup>. Therefore, MHD stresses are probably the dominant factor in reviving the stalled shock in rapidly rotating progenitors. Furthermore, we observe the formation of large-scale, structured toroidal magnetic field near the rotation axis of the protoneutron star in the later stages of the 50-m simulation (Fig. 4c and Extended Data Fig. 3d). This large-scale field is not present in the initial data (Fig. 4a), nor does it develop in the lower-resolution cases (Fig. 4b and Extended Data Fig. 3a–c). This magnetar-strength toroidal field close to the rotation axis is a strong indication that hoop stresses, which favour the formation of MHD-powered outflows, are present along the poles<sup>7,27</sup>. Velocity vectors along the rotation axis are pointing outwards towards the end

of the 50-m simulation, indicating the successful formation of bipolar outflows. (Extended Data Fig. 4).

Our findings have implications for stellar collapse in rapidly rotating massive stars. The MRI is a weak-field instability (that is, its growth time,  $\tau_{\text{MRI}}$ , does not depend on the strength of the magnetic field), and the observed rapid exponential-folding time of  $\tau \approx 0.5$  ms is short enough that the scenario presented here is viable even for much weaker initial seed fields. In addition, the MRI has been shown to operate efficiently in purely toroidal, mixed poloidal/toroidal, and random magnetic-field configurations<sup>2</sup>. Hence, we expect our results to hold for arbitrary precollapse magnetic-field configurations. Moreover, low-order multipole  $m = 1$  instabilities, shown to alter the explosion geometry of jet explosions in the full three-dimensional simulations of ref. 9, will start to become relevant only after a large-scale toroidal field of magnetar strength has been built up (the instability criterion depends on having an ultrastrong toroidal field present in the first place<sup>9</sup>). This makes MHD-driven explosions a likely scenario in rapidly rotating progenitors independently of the initial magnetization of the star, with the explosion geometry probably being of the double-lobe form shown in ref. 9. Finally, the large-scale build-up of magnetic field in the shear layer of the protoneutron star demonstrates that MRI-driven turbulence is a promising mechanism for forming pulsars and magnetars in rapidly rotating stellar collapse. This indicates that rapidly rotating massive stars can also account for potentially magnetar-powered superluminous supernovae<sup>25,26</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 7 April; accepted 23 September 2015.**

**Published online 30 November 2015.**

1. Fricke, K. Stability of rotating stars II. The influence of toroidal and poloidal magnetic fields. *Astron. Astrophys.* **1**, 388–398 (1969).
2. Balbus, S. A. & Hawley, J. F. A powerful local shear instability in weakly magnetized disks. I—linear analysis. II—nonlinear evolution. *Astrophys. J.* **376**, 214–233 (1991).
3. Frisch, U., Pouquet, A., Léorat, J. & Mazure, A. Possibility of an inverse cascade of magnetic helicity in magnetohydrodynamic turbulence. *J. Fluid Mech.* **68**, 769–778 (1975).
4. Moffatt, H. K. *Magnetic Field Generation in Electrically Conducting Fluids*. (Cambridge Univ. Press, 1978).
5. Brandenburg, A. & Subramanian, K. Astrophysical magnetic fields and nonlinear dynamo theory. *Phys. Rep.* **417**, 1–209 (2005).
6. Brandenburg, A. The inverse cascade and nonlinear alpha-effect in simulations of isotropic helical hydromagnetic turbulence. *Astrophys. J.* **550**, 824–840 (2001).
7. LeBlanc, J. M. & Wilson, J. R. A numerical example of the collapse of a rotating magnetized star. *Astrophys. J.* **161**, 541–551 (1970).

8. Burrows, A., Dessart, L., Livne, E., Ott, C. D. & Murphy, J. Simulations of magnetically driven supernova and hypernova explosions in the context of rapid rotation. *Astrophys. J.* **664**, 416–434 (2007).
9. Mösta, P. *et al.* Magnetorotational core-collapse supernovae in three dimensions. *Astrophys. J. Lett.* **785**, L29 (2014).
10. Takiwaki, T. & Kotake, K. Gravitational wave signatures of magnetohydrodynamically driven core-collapse supernova explosions. *Astrophys. J.* **743**, 30 (2011).
11. Akiyama, S., Wheeler, J. C., Meier, D. L. & Lichtenstadt, I. The magnetorotational instability in core-collapse supernova explosions. *Astrophys. J.* **584**, 954–970 (2003).
12. Thompson, T. A., Quataert, E. & Burrows, A. Viscosity and rotation in core-collapse supernovae. *Astrophys. J.* **620**, 861–877 (2005).
13. Drout, M. R. *et al.* The first systematic study of type Ibc supernova multi-band light curves. *Astrophys. J.* **741**, 97 (2011).
14. Soderberg, A. M. *et al.* Relativistic ejecta from X-ray flash XRF 060218 and the rate of cosmic explosions. *Nature* **442**, 1014–1017 (2006).
15. Hjorth, J. & Bloom, J. S. in *Gamma-Ray Bursts* (eds Wijers, R. A. M. J. & Woosley, S.) Ch. 9 (Cambridge Univ. Press, 2012).
16. Modjaz, M. Stellar forensics with the supernova-GRB connection. *Astron. Nachr.* **332**, 434–447 (2011).
17. Obergaulinger, M., Cerdá-Durán, P., Müller, E. & Aloy, M. A. Semi-global simulations of the magneto-rotational instability in core collapse supernovae. *Astron. Astrophys.* **498**, 241–271 (2009).
18. Masada, Y., Takiwaki, T. & Kotake, K. Magnetohydrodynamic turbulence powered by magnetorotational instability in nascent protoneutron stars. *Astrophys. J. Lett.* **798**, L22 (2015).
19. Guilet, J., Müller, E. & Janka, H.-T. Neutrino viscosity and drag: impact on the magnetorotational instability in protoneutron stars. *Mon. Not. R. Astron. Soc.* **447**, 3992–4003 (2015).
20. Sawai, H., Yamada, S. & Suzuki, H. Global simulations of magnetorotational instability in the collapsed core of a massive star. *Astrophys. J. Lett.* **770**, L19 (2013).
21. Galama, T. J. *et al.* An unusual supernova in the error box of the  $\gamma$ -ray burst of 25 April 1998. *Nature* **395**, 670–672 (1998).
22. Woosley, S. E. & Heger, A. The progenitor stars of gamma-ray bursts. *Astrophys. J.* **637**, 914–921 (2006).
23. Thompson, C. & Duncan, R. C. Neutron star dynamos and the origins of pulsar magnetism. *Astrophys. J.* **408**, 194–217 (1993).
24. Duncan, R. C. & Thompson, C. Formation of very strongly magnetized neutron stars—implications for gamma-ray bursts. *Astrophys. J. Lett.* **392**, L9–L13 (1992).
25. Nicholl, M. *et al.* Slowly fading super-luminous supernovae that are not pair-instability explosions. *Nature* **502**, 346–349 (2013).
26. Greiner, J. *et al.* A very luminous magnetar-powered supernova associated with an ultra-long  $\gamma$ -ray burst. *Nature* **523**, 189–192 (2015).
27. Wheeler, J. C., Meier, D. L. & Wilson, J. R. Asymmetric supernovae from magnetocentrifugal jets. *Astrophys. J.* **568**, 807–819 (2002).
28. Pessah, M. E. & Goodman, J. On the saturation of the magnetorotational instability via parasitic modes. *Astrophys. J. Lett.* **698**, L72–L76 (2009).
29. Goodman, J. & Xu, G. Parasitic instabilities in magnetized, differentially rotating disks. *Astrophys. J.* **432**, 213–223 (1994).
30. Ott, C. D., Burrows, A., Thompson, T. A., Livne, E. & Walder, R. The spin periods and rotational profiles of neutron stars at birth. *Astrophys. J. (Suppl.)* **164**, 130–155 (2006).

**Acknowledgements** We thank S. Couch, J. Zrake, D. Tsang, C. Wheeler, E. Bentivegna and I. Hinder for discussions. This research was supported by National Science Foundation (NSF) grants AST-1212170, PHY-1151197 and OCI-0905046; by NASA through the Einstein Fellowship Program, grants PF5-160140 (to P.M.) and PF3-140114 (to L.F.R.); by a National Science and Engineering Research Council of Canada (NSERC) award to E.S.; and by the Sherman Fairchild Foundation. The simulations were carried out on the NSF/National Center for Supercomputing Applications (NCSA) BlueWaters supercomputer (PRAC ACI-1440083).

**Author Contributions** P.M. contributed to project planning and leadership, simulation code development, simulations, simulation analysis, visualization, interpretation of results and manuscript preparation. C.D.O. led the group, conceived the idea for the project, and contributed to project planning and leadership, interpretation and manuscript preparation. D.R. contributed to simulation analysis, interpretation, simulation code development and manuscript preparation. L.F.R. interpreted the results and reviewed the manuscript. E.S. contributed to simulation code development and manuscript review. R.H. contributed to development of the simulation code and visualization software, and reviewed the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.M. ([pmoesta@berkeley.edu](mailto:pmoesta@berkeley.edu)).

## METHODS

**Initial conditions—stellar collapse simulation.** We start by performing a dynamical space-time ideal MHD simulation with adaptive mesh refinement (AMR) of the  $25M_{\odot}$  (at zero-age-main-sequence) presupernova model E25 (ref. 31), with initial conditions for differential rotation as in ref. 9 (initial central angular velocity of the iron core is  $2.8 \text{ rad s}^{-1}$ ;  $x_0 = 500 \text{ km}$ ;  $z_0 = 2,000 \text{ km}$ ;  $M_{\odot}$ , mass of the Sun). This model could be considered as a type Ic-bl/hypernova and long  $\gamma$ -ray burst progenitor<sup>22</sup>. At the onset of collapse, we set up a modified dipolar magnetic field structure from a vector potential  $A$  with components  $A_r = A_{\theta} = 0$  and  $A_{\phi} = B_0(r_0^3)(r^3 + r_0^3)^{-1} r \sin\theta$ , where  $r$  is the radius,  $r_0 = 1,000 \text{ km}$  (as in ref. 9) is a parameter controlling the fall-off of the magnetic field, and  $B_0 = 10^{10} \text{ G}$  sets the initial strength of the magnetic field. This progenitor seed field is not unreasonable for  $\gamma$ -ray-burst supernova progenitor cores<sup>22,32</sup>. With the grid set-up (nine levels of box-in-box AMR; finest resolution  $dx = 375 \text{ m}$ ) and methods being identical to those in refs 9 and 33, we follow this simulation until  $t_{\text{map}} = 20 \text{ ms}$  after core bounce. At this time, the initial supernova shockwave has stalled at a radius of about  $130 \text{ km}$ .

Extended Data Figs 5 and 6 show the radial profiles of important state variables (density, entropy, angular velocity and fast magnetosonic speed) of the simulation at the time of mapping. Both the protoneutron star and the postshock region have reached a quasi-equilibrium state, and the underlying space-time changes only very slowly and secularly, allowing us to carry out subsequent high-resolution GRMHD simulations that assume a fixed background space-time for about  $10$ – $20 \text{ ms}$ . The resolution of the AMR box covering the shear layer of the protoneutron star in this initial simulation is  $dx = 750 \text{ m}$ , but to resolve the FGM of the MRI for the chosen initial magnetic field of  $10^{10} \text{ G}$ , a linear resolution of at least  $dx = 100 \text{ m}$  is required<sup>34</sup>. This is why the common method of obtaining the field strength necessary to power a magnetorotational explosion ( $\geq 10^{15} \text{ G}$ ) has been by flux compression ( $B \propto \rho^{2/3}$ ; amplification by a factor of about  $10^3$ ;  $\rho$  is the density of the gas in the collapsing core) from unrealistically high seed fields ( $B \geq 10^{12} \text{ G}$  precollapse)<sup>8,10,35,36</sup>.

**Background flow stability analysis.** A magnetized fluid is unstable to weak-field shearing modes in the presence of a negative angular velocity gradient that is not compensated for by compositional or entropy gradients of the fluid<sup>2</sup>. At the time of mapping of the initial AMR simulation to the high-resolution domain, the plasma in the shocked region around the protoneutron star is locally unstable to weak-field shearing modes, as given by  $C_{\text{MRI}} \equiv (\omega_{\text{BV}}^2 + r \times d\Omega^2/dr)/\Omega^2 < 0$  (refs 2, 34, 37). Here,  $C_{\text{MRI}}$  is the stability criterion of the MRI;  $\omega_{\text{BV}}$  is the Brunt–Väisälä frequency indicating convective stability/instability;  $r \frac{d\Omega^2}{dr}$  characterizes the rotational shear; and  $\Omega$  is the angular velocity. We follow refs 11 and 37 and calculate the stability criterion  $C_{\text{MRI}}$ , as well as the wavelength ( $\lambda_{\text{FGM}}$ ) and growth time ( $\tau_{\text{FGM}}$ ) of the FGM of the MRI, in two-dimensional  $x$ – $y$  and  $x$ – $z$  slices through our three-dimensional domain.

To better approximate the background flow in our three-dimensional AMR stellar collapse simulation, we average in space and time. We first carry out a spatial averaging step and calculate averaged versions of the state variables of our simulation (for example, the spatially averaged density  $\rho_i$ ) at every time step. For that, we choose a centred stencil that takes into account three points in each direction (this is the maximum number of points that we have available at AMR component boundaries). Because this is insufficient to get a large enough sample of points for the averaging procedure, we also calculate a moving time average of the form  $\rho_{\text{av},i} = \alpha \rho_i + (1 - \alpha) \rho_{\text{av},i-1}$ , where  $i$  denotes the current time step and  $i-1$  the previous one. We choose a weight function for each data set in the moving average as  $\alpha = 2(n \Delta t / \Delta t_{\text{coarse}} + 1)^{-1}$ , where  $\Delta t$  is the time step on the current refinement level, and  $\Delta t_{\text{coarse}}$  is the time step of the coarsest level. This choice of weight function guarantees that 86% of the data in the average comprise the last  $n$  time-step data sets. The time-step size in our AMR simulation on the refinement level that contains the shear layer around the protoneutron star is  $\Delta t = 5 \times 10^{-4} \text{ ms}$ , and we choose  $n$  such that  $\alpha = 2,000$ , ensuring temporal averaging over a timescale of about  $1 \text{ ms}$ . We calculate  $C_{\text{MRI}}$ ,  $\lambda_{\text{FGM}}$  and  $\tau_{\text{FGM}}$  from the space and time averages of the state variables in our simulation (Extended Data Fig. 2).

**Mapping to a high-resolution computational domain.** Next, we map the configuration to a three-dimensional domain with uniform spacing of the form  $x, y, z = [-66.5 \text{ km}, 66.5 \text{ km}]$  for four resolutions,  $h = [500 \text{ m}, 200 \text{ m}, 100 \text{ m}, 50 \text{ m}]$ . To guarantee divergence-free initial data for the magnetic field, we carry out a constraint projection step after we have interpolated the magnetic field to the new domain. This is technically challenging as we have to make sure that all operators used in the projection are consistent in their definition with the discrete form of the divergence operator maintained in our specific implementation of constrained transport<sup>33</sup>. We use a discrete analogue of the Helmholtz decomposition<sup>36</sup> to decompose the magnetic field into a discrete curl,  $\nabla_h \times$ , and a discrete gradient,  $\nabla_h$ :

$$\mathbf{B} = \nabla_h \times \mathbf{A} + \nabla_h \Phi \quad (1)$$

where  $\Phi$  is a discrete scalar field. The discrete divergence,  $\nabla_h \cdot$ , of equation (1) leads to a discrete Poisson equation:

$$\nabla_h \cdot \mathbf{B} = \Delta_h \Phi \quad (2)$$

where  $\Delta_h$  is the discrete Laplace operator. We solve equation (2) augmented with homogeneous Dirichlet boundary conditions to machine precision for  $\Phi$  using the conjugate gradient solver provided by the PETSc<sup>2</sup> library in combination with the parallel algebraic multigrid preconditioner HYPRE<sup>37</sup>. We then obtain a divergence-free field,  $\mathbf{B}'$ , from the projection  $\mathbf{B}' = \mathbf{B} - \nabla_h \Phi$ . Finally, we recompute  $\nabla_h \cdot \mathbf{B}'$  to check that it is zero to floating-point precision.

**High-resolution turbulence simulations.** We perform ideal, fixed background space-time, GRMHD simulations using the open-source Einstein Toolkit<sup>33,38</sup> with WENO5 reconstruction<sup>39,40</sup>, the HLLC Riemann solver<sup>41</sup> and constrained transport<sup>42</sup> for maintaining  $\nabla \cdot \mathbf{B} = 0$ . We use the  $K_0 = 220 \text{ MeV}$  variant of the finite-temperature nuclear equation of state of ref. 43, and the neutrino leakage/heating approximations described in refs 44, 45, with a heating scale factor  $f_{\text{heat}} = 1.0$ . We perform simulations on a domain with uniform spacing of the form  $x, y = [0 \text{ km}, 66.5 \text{ km}]$  and  $z = [-66.5 \text{ km}, 66.5 \text{ km}]$  for four resolutions,  $h = [500 \text{ m}, 200 \text{ m}, 100 \text{ m}, 50 \text{ m}]$ , in quadrant symmetry three dimensions ( $90^\circ$  rotational symmetry in the  $x$ – $y$  plane). We keep all variables at the boundary fixed in time. This is justifiable for several reasons. First, the accretion boundary flow itself only changes on timescales longer than those simulated. Second, the fast magnetosonic speed (Extended Data Figs 5d and 6c) is of the order of a few per cent of the speed of light throughout the high-resolution computational domain. This implies a boundary crossing time for the simulation box of about  $20 \text{ ms}$ . This leaves the results in the shear layer unaltered by boundary effects for the simulated times of  $10 \text{ ms}$ . Additionally, as the cylindrically rotating flow in the shocked region is rotating in and out of the purely Cartesian boundary zones, sound waves can be reflected at the boundaries. Although these reflections are not necessarily unrealistic, as there will be perturbations in the shocked region of any rotating iron core, they pose an additional complication for the numerical stability of the simulations<sup>46</sup>. We find these reflections to be minimal in the hydrodynamical variables themselves, but they do cause spurious oscillations in the magnetic field towards the boundary zones. To prevent these oscillations at the outer boundary, without affecting the solution in the shear layer around the protoneutron star, we apply diffusivity at the level of the induction equation for the magnetic field via a modified Ohm's law. We choose  $\mathbf{E} = -\mathbf{v} \times \mathbf{B} + \eta \mathbf{J}$ , where  $\mathbf{J} = \nabla \times \mathbf{B}$  is the three-current density; we set  $\eta = \eta_0(0.5 + 0.5 \tanh((r - r_{\text{diff}})b^{-1}))$  with  $\eta_0 = 10^{-2}$ ,  $r_{\text{diff}} = 40 \text{ km}$  and  $b = 3 \text{ km}$ . That is, we apply diffusivity only in a region outside of radius  $r_{\text{diff}}$  and transition smoothly over a blending zone with width  $b$  to no diffusivity inside  $r_{\text{diff}}$ .

**Turbulent kinetic and magnetic energy spectra.** We compute spectra of the turbulent kinetic and magnetic energy as instantaneous snapshots using the

discrete Fourier transform  $\hat{\mathbf{u}}(\mathbf{k}) = \sum_{\mathbf{x}} \mathbf{u}(\mathbf{x}) \exp\left(-2\pi i \frac{\mathbf{k} \cdot \mathbf{x}}{L}\right) \left(\frac{L}{N}\right)^3$  (ref. 47), where  $\mathbf{u}$

is a vector field,  $L$  is the extent of the computational box, and  $N$  the number of grid points in the computational box. The spectra shown in Fig. 3 are densitized to better reflect the overall energy contained in the turbulent kinetic motion and the magnetic field. We show the spectra of the non-densitized turbulent velocity in Extended Data Fig. 7a, and the non-densitized magnetic field in Extended Data Fig. 7b, and also window the data to account for the non-periodicity at the boundaries of our computational domain. For that, we use a mollifier of the form  $m(x) = \exp\left\{1 - \left[1 - \left(\frac{x - d}{d}\right)^2\right]^{-1}\right\}$ , and respectively for  $y$  and  $z$ . This effectively

blends the data to zero over a stencil width  $d$  at the outer boundary. We choose  $d = 3$ , but note that other choices yield similar results. These non-densitized and windowed spectra illustrate that the lack of an exponential turnoff at large  $k$  in the turbulent kinetic energy in Fig. 3 is due to the inclusion of the nearly discontinuous density fall-off at the edge of the protoneutron star core (at  $r \approx 12 \text{ km}$ ) in the calculation of the spectrum for Fig. 3 and the non-periodicity of our computational domain. The non-densitized and windowed turbulent kinetic energy spectrum in Extended Data Fig. 7 is compensated for  $k^{-5/3}$  scaling (as expected according to Kolmogorov theory<sup>48</sup>). We observe a slightly steeper scaling between  $k^{-5/3}$  and  $k^{-2}$ . Within the first  $3 \text{ ms}$ , there is a rapid transition into a fully turbulent state at large  $k$  (Fig. 3b and Extended Data Fig. 7a). Afterwards, the turbulent kinetic energy decreases at large  $k$  and the spectrum gradually evolves towards a steeper fall-off. There is no increase in the turbulent kinetic energy at small values of  $k$  at late times. The magnetic energy, similarly to the turbulent kinetic energy, peaks at large  $k$  at  $t - t_{\text{map}} \approx 3 \text{ ms}$ , which correlates well with the observed saturation of the maximum toroidal field shown in Fig. 1. Subsequently, the magnetic energy at small  $k$  grows



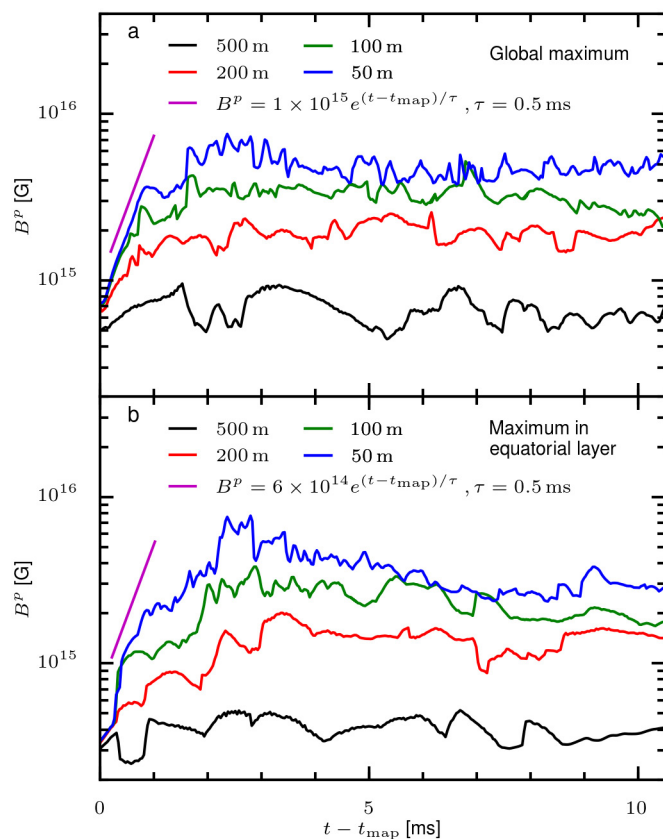
first exponentially and then linearly with time. This picture is consistent with energy being extracted from the turbulent kinetic motion at large  $k$  and being pumped into an inverse cascade that leads to growth of magnetic-field energy at small values of  $k$ . As the kinematic phase ends and transitions into saturation, magnetic fields and numerical resistivity become important for the evolution<sup>49</sup>. This may explain the transition to linear growth. We also observe a superposed 2-ms modulation on top of the  $k=4$  exponential growth that corresponds roughly to the Alfvén crossing time across the shear layer ( $t_{A, \text{shear}} \approx 2$  ms).

**Angle-averaged magnetic flux and poloidal current.** We compute the two-dimensional angle-averaged (in  $\varphi$ ) magnetic flux and poloidal current to determine which magnetic-field structures are global in  $\varphi$  (Extended Data Figs 3 and 4). The magnetic flux is computed as  $\int_0^{\omega_{\text{max}}} \omega B^z d\omega$  and the current as  $\mathbf{J} = \nabla \times \mathbf{B}$ . The isocontours of the magnetic flux represent the poloidal field lines, while the poloidal current approximates the toroidal magnetic field. We find that the shear layer of the protoneutron star distorts the initial poloidal magnetic field of the iron core, but we find no emerging global poloidal field created from turbulence. The toroidal field (poloidal current), however, does show a global structure that roughly fills the width of the shear layer in the polar region of our simulation, supporting the idea that the toroidal magnetar-strength field in our simulations (see also Fig. 4) truly is global in  $\varphi$ .

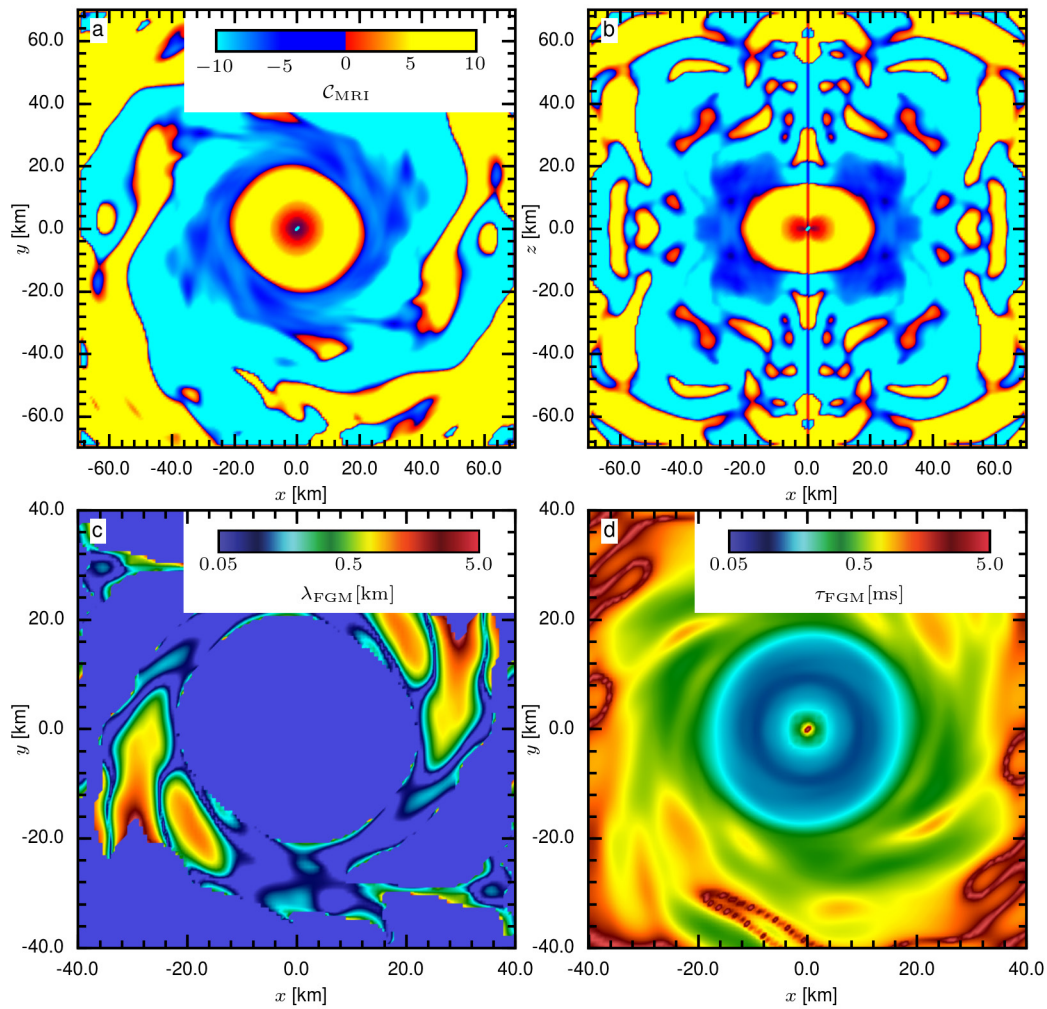
**Limitations of this study.** The limitations of this study are finite resolution of the simulations (most visible in the not-fully-converged saturation magnetic field), and the sensitivity of the detailed turbulent state to the numerical methods. Also, the impact of the imposed 90° rotational symmetry has to be investigated. Ultimately, high-resolution simulations such as these have to be embedded back into a full-star simulation to determine the detailed shock revival and explosion geometry.

**Code availability.** All computer code used here that is not already freely available, and the initial data, are available at <http://stellarcollapse.org>.

31. Heger, A. & Langer, N. Presupernova evolution of rotating massive stars. II. Evolution of the surface properties. *Astrophys. J.* **544**, 1016–1035 (2000).
32. Wheeler, J. C., Kagan, D. & Chatzopoulos, E. The role of the magnetorotational instability in massive stars. *Astrophys. J.* **799**, 85 (2015).
33. Mösta, P. *et al.* GRHydro: a new open-source general-relativistic magnetohydrodynamics code for the Einstein toolkit. *Class. Quantum Gravity* **31**, 015005 (2014).
34. Obergaulinger, M., Aloy, M. A. & Müller, E. Axisymmetric simulations of magneto-rotational core collapse: dynamics and gravitational wave signal. *Astron. Astrophys.* **450**, 1107–1134 (2006).
35. Winteler, C. *et al.* Magnetorotationally driven supernovae as the origin of early galaxy  $r$ -process elements? *Astrophys. J. Lett.* **750**, L22 (2012).
36. Nishimura, N., Takiwaki, T. & Thielemann, F.-K. The  $r$ -process nucleosynthesis in the various jet-like explosions of magnetorotational core-collapse supernovae. *Astrophys. J.* **810**, 109 (2015).
37. Balbus, S. A. & Hawley, J. F. Instability, turbulence, and enhanced transport in accretion disks. *Rev. Mod. Phys.* **70**, 1–53 (1998).
38. Löffler, F. *et al.* The Einstein toolkit: a community computational infrastructure for relativistic astrophysics. *Class. Quantum Gravity* **29**, 115001 (2012).
39. Reisswig, C. *et al.* Three-dimensional general-relativistic hydrodynamic simulations of binary neutron star coalescence and stellar collapse with multipatch grids. *Phys. Rev. D* **87**, 064023 (2013).
40. Tchekhovskoy, A., McKinney, J. C. & Narayan, R. WHAM: a WENO-based general relativistic numerical scheme. I. Hydrodynamics. *Mon. Not. R. Astron. Soc.* **379**, 469–497 (2007).
41. Einfeldt, B. in *Shock Tubes and Waves* (ed. Groenig, H.) 671–676 (VCH, 1988).
42. Tóth, G. The  $\nabla \cdot \mathbf{B} = 0$  constraint in shock-capturing magnetohydrodynamics codes. *J. Comput. Phys.* **161**, 605–652 (2000).
43. Lattimer, J. M. & Douglas Swesty, F. A generalized equation of state for hot, dense matter. *Nucl. Phys. A* **535**, 331–376 (1991).
44. O'Connor, E. & Ott, C. D. A new open-source code for spherically symmetric stellar collapse to neutron stars and black holes. *Class. Quantum Gravity* **27**, 4103 (2010).
45. Ott, C. D. *et al.* Correlated gravitational wave and neutrino signals from general-relativistic rapidly rotating iron core collapse. *Phys. Rev. D* **86**, 024026 (2012).
46. Bogovalov, S. V. Boundary conditions and critical surfaces in astrophysical MHD winds. *Astron. Astrophys.* **323**, 634–643 (1997).
47. Eswaran, V. & Pope, S. B. An examination of forcing in direct numerical simulations of turbulence. *Comput. Fluids* **16**, 257–278 (1988).
48. Frisch, U. *Turbulence. The Legacy of A. N. Kolmogorov* (Cambridge Univ. Press, 1995).
49. Zrake, J. & MacFadyen, A. I. Magnetic energy production by turbulence in binary neutron star mergers. *Astrophys. J. Lett.* **769**, L29 (2013).



**Extended Data Figure 1 | Evolution of the maximum poloidal magnetic field.** Both panels show the maximum poloidal magnetic field,  $B^p$ , as a function of time for the four resolutions: 500 m, 200 m, 100 m and 50 m. **a**, The global maximum field. **b**, The maximum field in a thin layer above and below the equatorial plane ( $-7.5 \text{ km} \leq z \leq 7.5 \text{ km}$ ). The purple line indicates exponential growth with an exponential-folding time,  $\tau_{\text{FGM}}$ , of 0.5 ms.

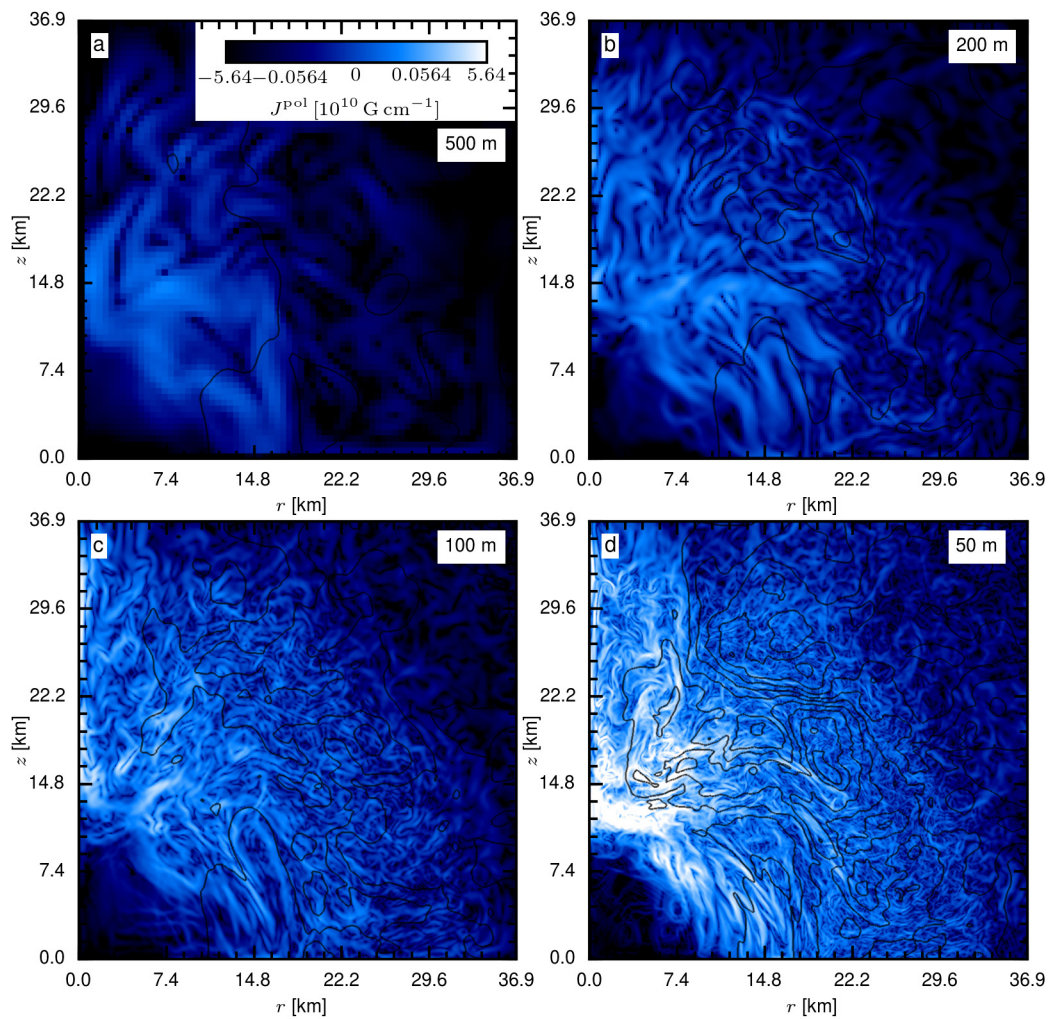


#### Extended Data Figure 2 | Background flow stability analysis.

**a, b,** The stability criterion  $C_{\text{MRI}}$  20 ms after core bounce for the initial stellar collapse simulation. **a,** A two-dimensional  $x$ - $y$  slice ( $z=0$ ) through the three-dimensional domain; **b,** an  $x$ - $z$  slice ( $y=0$ ). Yellow and red

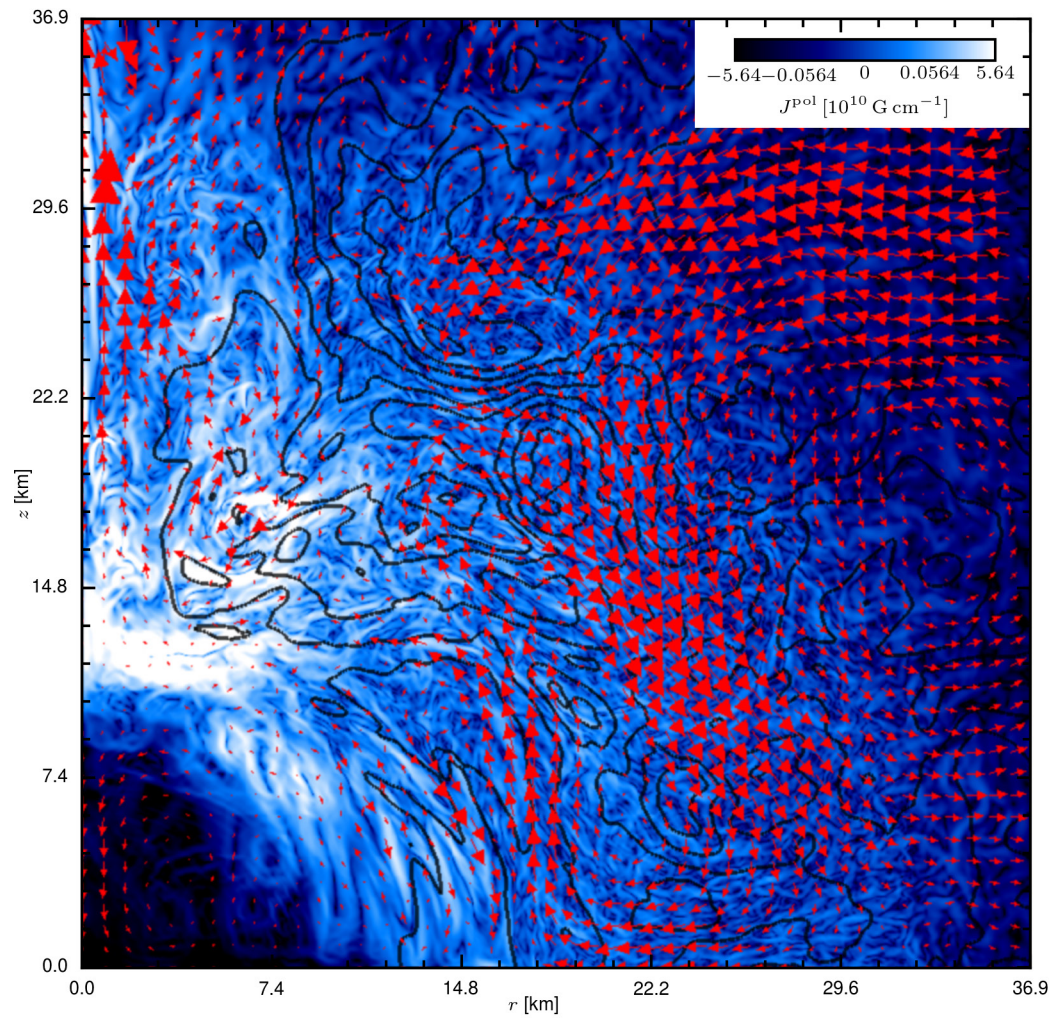
indicate regions that are stable to shearing modes; dark blue and light blue indicate unstable regions. **c,** The wavelength,  $\lambda_{\text{FGM}}$ , of the FGM of the MRI. **d,** The growth time of the FGM,  $\tau_{\text{FGM}}$ . Panels **c** and **d** are zoomed in on the shear layer around the protoneutron star.



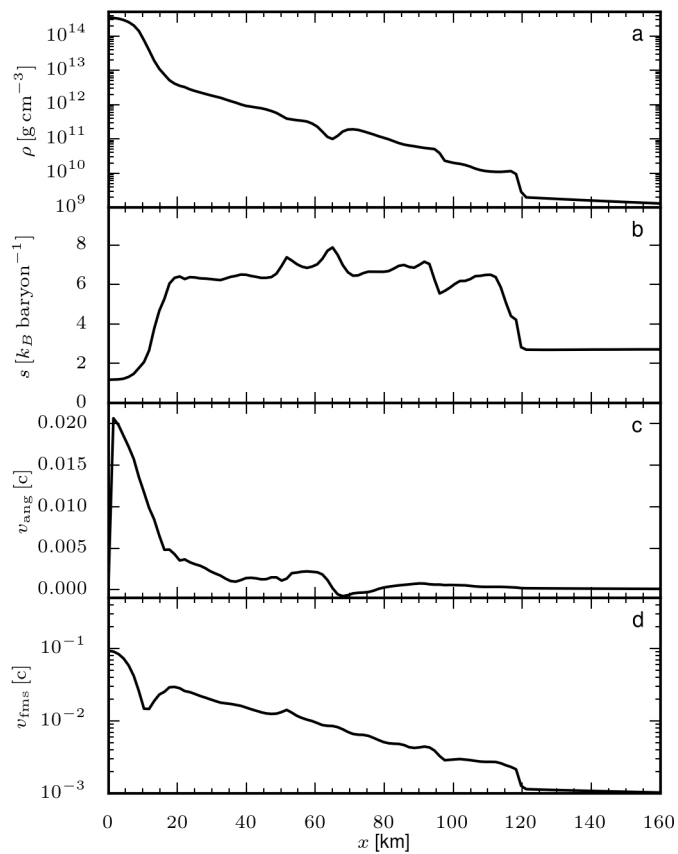


**Extended Data Figure 3 | Angle-averaged poloidal magnetic current and magnetic flux.** All panels show  $r$ - $z$  slices (cylindrical coordinates, angle-averaged in  $\varphi$ ) of the poloidal magnetic current ( $J^{\text{pol}}$ , colour-coded)

and superposed contours of magnetic flux (black lines) at  $t - t_{\text{map}} = 10.3 \text{ ms}$  (final simulated time). **a**, The 500-m simulation; **b**, the 200-m simulation; **c**, the 100-m simulation; **d**, the 50-m simulation.

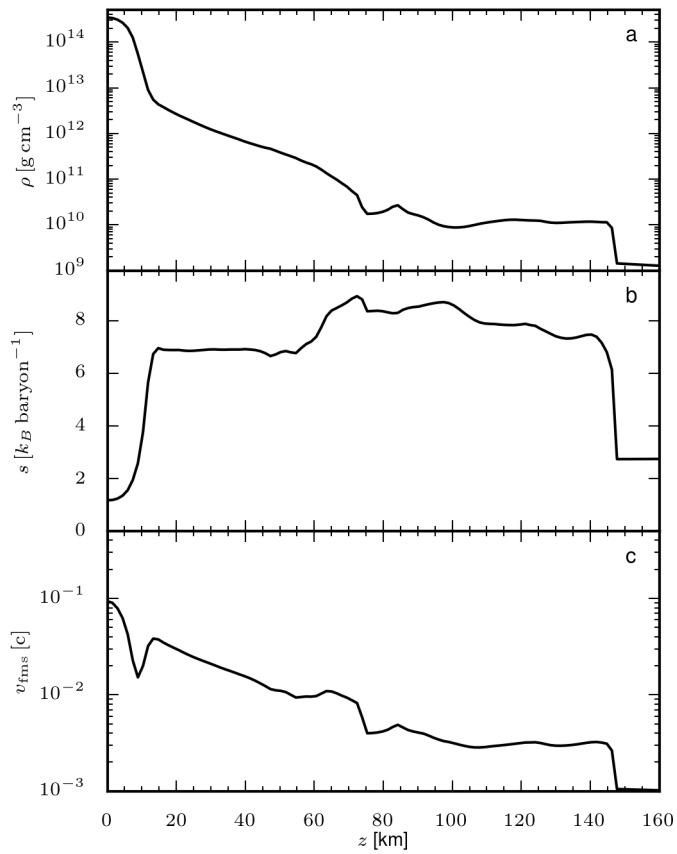


**Extended Data Figure 4 | Angle-averaged poloidal magnetic current and velocity vectors.** The figure shows  $r$ - $z$  slices (cylindrical coordinates, angle-averaged in  $\varphi$ ) of the poloidal magnetic current ( $J^{\text{pol}}$ , colour-coded) and superposed velocity vectors (red arrows) at  $t - t_{\text{map}} = 10.3 \text{ ms}$  (final simulated time).

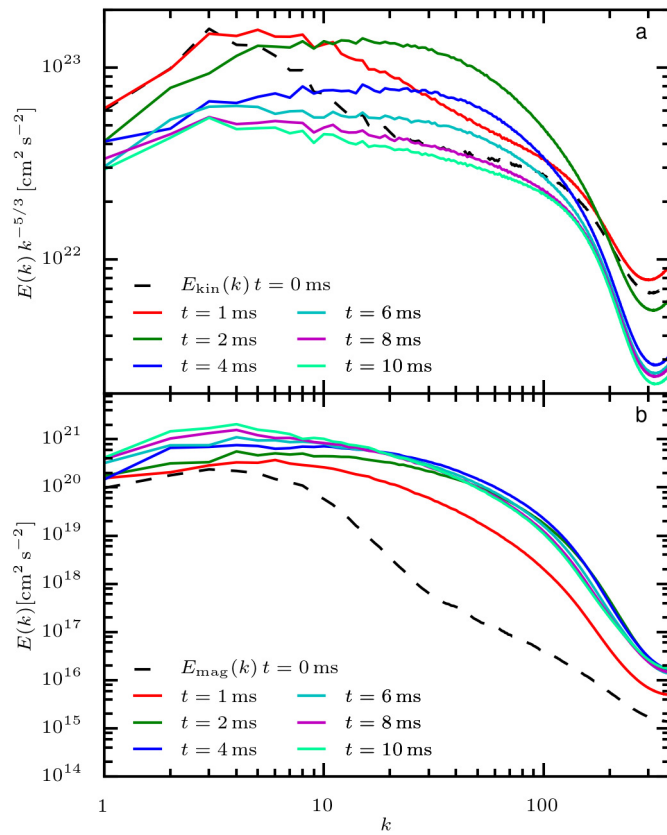


**Extended Data Figure 5 | AMR stellar collapse simulation.** All panels show profiles along the  $x$  direction of the initial stellar collapse simulation, 20 ms after core bounce. **a**, Density ( $\rho$ ); **b**, entropy ( $s$ ),  $k_B$  is the Boltzmann constant; **c**, angular velocity ( $v_{\text{ang}}$ ); **d**, fast magnetosonic speed ( $v_{\text{fms}}$ ).





**Extended Data Figure 6 | AMR stellar collapse simulation.** All panels show profiles along the  $z$  direction of the initial stellar collapse simulation, 20 ms after core bounce. **a**, Density; **b**, entropy; **c**, fast magnetosonic speed.



**Extended Data Figure 7 | Non-densitized turbulent kinetic and electromagnetic energy spectra.** **a**, A time series of non-densitized turbulent kinetic energy spectra,  $E_{\text{kin}}(k)$ , compensated for Kolmogorov scaling ( $k^{-5/3}$ ), as a function of the dimensionless wavenumber  $k$ . **b**, A time series of non-densitized magnetic energy spectra,  $E_{\text{mag}}(k)$ , as a function of the dimensionless wavenumber  $k$ . In both panels, the initial spectrum at  $t - t_{\text{map}} = 0$  ms (dashed black line) is shown for reference.

# Multi-element logic gates for trapped-ion qubits

T. R. Tan<sup>1</sup>, J. P. Gaebler<sup>1</sup>, Y. Lin<sup>1†</sup>, Y. Wan<sup>1</sup>, R. Bowler<sup>1†</sup>, D. Leibfried<sup>1</sup> & D. J. Wineland<sup>1</sup>

Precision control over hybrid physical systems at the quantum level is important for the realization of many quantum-based technologies. In the field of quantum information processing (QIP) and quantum networking, various proposals discuss the possibility of hybrid architectures<sup>1</sup> where specific tasks are delegated to the most suitable subsystem. For example, in quantum networks, it may be advantageous to transfer information from a subsystem that has good memory properties to another subsystem that is more efficient at transporting information between nodes in the network. For trapped ions, a hybrid system formed of different species introduces extra degrees of freedom that can be exploited to expand and refine the control of the system. Ions of different elements have previously been used in QIP experiments for sympathetic cooling<sup>2</sup>, creation of entanglement through dissipation<sup>3</sup>, and quantum non-demolition measurement of one species with another<sup>4</sup>. Here we demonstrate an entangling quantum gate between ions of different elements which can serve as an important building block of QIP, quantum networking, precision spectroscopy, metrology, and quantum simulation. A geometric phase gate between a  ${}^9\text{Be}^+$  ion and a  ${}^{25}\text{Mg}^+$  ion is realized through an effective spin–spin interaction generated by state-dependent forces induced with laser beams<sup>5–9</sup>. Combined with single-qubit gates and same-species entangling gates, this mixed-element entangling gate provides a complete set of gates over such a hybrid system for universal QIP<sup>10–12</sup>. Using a sequence of such gates, we demonstrate a CNOT (controlled-NOT) gate and a SWAP gate<sup>13</sup>. We further demonstrate the robustness of these gates against thermal excitation and show improved detection in quantum logic spectroscopy<sup>14</sup>. We also observe a strong violation of a CHSH (Clauser–Horne–Shimony–Holt)-type Bell inequality<sup>15</sup> on entangled states composed of different ion species.

Trapped ions of different elements vary in mass, internal atomic structure, and spectral properties, features that can make certain species suited for particular tasks such as storing quantum information, high-fidelity readout, fast logic gates, or interfacing between local processors and photon interconnects. One important advantage of a hybrid system incorporating trapped ions of different elements is the ability to manipulate and measure one type of qubit using laser beams with negligible effects on the other since the resonant transition wavelengths differ substantially. When scaling trapped-ion systems to greater numbers and density of ions, it will be advantageous to perform fluorescence detection on individual qubits without inducing decoherence on neighbouring qubits due to uncontrolled photon scattering. To provide this function in a hybrid system one can use an entangling gate to transfer the qubit states to another ion species which is then detected without perturbing the qubits. This readout protocol could be further generalized to error correction schemes by extracting the error syndromes to the readout species while the computational qubits remain in the code. Another application could be in building photon interconnects between trapped-ion devices. Here, one species may be better suited for memory while the other is more favourable for coupling to photons<sup>16,17</sup>.

A mixed-element gate can also improve the readout in quantum logic spectroscopy (QLS)<sup>14</sup>. In conventional quantum logic readout,

the state of the clock or qubit ion is transferred to a motional state and in turn transferred to the detection ion, which is then detected with state-dependent fluorescence. In this case, the transfer fidelity directly depends on the purity of the motional state. In contrast, transfer using the gate discussed here can be insensitive to the motion, as long as the ions are in the Lamb–Dicke regime<sup>18</sup>. This advantage extends to entanglement-assisted quantum non-demolition (QND) readout of qubit or clock ions, which can lower the overhead in time and number of readout ions as the number of clock ions increases<sup>19</sup>.

In our experiment, we use a beryllium ( ${}^9\text{Be}^+$ ) ion and a magnesium ( ${}^{25}\text{Mg}^+$ ) ion separated by approximately  $4\,\mu\text{m}$  along the axis of a linear Paul trap<sup>20</sup>. The addressing lasers for each ion (wavelength  $\lambda \approx 313\,\text{nm}$  for  ${}^9\text{Be}^+$  and  $\lambda \approx 280\,\text{nm}$  for  ${}^{25}\text{Mg}^+$ ) illuminate both ions. The qubits are encoded in hyperfine states of the ions. We choose  $|F=2, m_F=0\rangle = |\downarrow\rangle_{\text{Be}}$  and  $|1, 1\rangle = |\uparrow\rangle_{\text{Be}}$  as the  ${}^9\text{Be}^+$  qubit states, and  $|2, 0\rangle = |\downarrow\rangle_{\text{Mg}}$  and  $|3, 1\rangle = |\uparrow\rangle_{\text{Mg}}$  for the  ${}^{25}\text{Mg}^+$  qubit. The Coulomb coupling between the ions gives rise to two shared motional normal modes along the trap axis. A magnetic field of  $11.945\,\text{mT}$  is applied at  $45^\circ$  with respect to the trap axis. At this field, the  ${}^9\text{Be}^+$  qubit transition frequency is, to first order, insensitive to external magnetic field fluctuations<sup>21</sup>. The magnetic field sensitivity of the  ${}^{25}\text{Mg}^+$  qubit is approximately  $430\,\text{kHz}\,\text{mT}^{-1}$ . By measuring the decay of Ramsey interference fringes versus time between the Ramsey pulses on each qubit transition, we determine the  ${}^9\text{Be}^+$  qubit's coherence time to be approximately  $1.5\,\text{s}$ . The  ${}^{25}\text{Mg}^+$  qubit coherence time is approximately  $6\,\text{ms}$ , limited by magnetic field fluctuations. We verified that the phase and contrast of Ramsey experiments on one species do not change measurably in the presence of light addressing the other species. This shows that the spectral separation is sufficient to isolate the species.

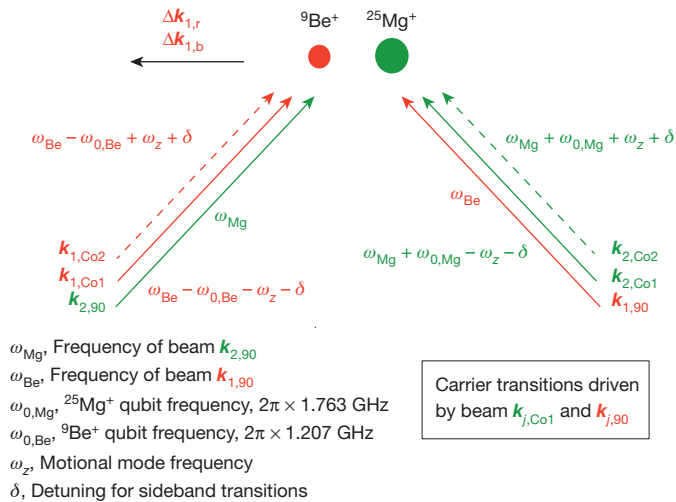
Entanglement between the two ions is achieved through a Mølmer–Sørensen (MS) spin–spin interaction<sup>5–8</sup> induced by laser-driven stimulated Raman transitions<sup>18</sup>. Starting in the state  $|\uparrow\rangle_{\text{Be}}|\uparrow\rangle_{\text{Mg}} = |\uparrow\uparrow\rangle$ , the interaction can produce the Bell state  $\Phi_+ = \frac{1}{\sqrt{2}}(|\downarrow\downarrow\rangle + |\uparrow\uparrow\rangle)$  (see Methods).

The laser beam configurations to induce coherent Raman transitions are analogous for each element; for brevity, we will only describe the configuration for  ${}^9\text{Be}^+$  (red in Fig. 1). Three laser beams, labelled by their wave vectors  $\mathbf{k}_{1,\text{Co}1}$ ,  $\mathbf{k}_{1,\text{Co}2}$ , and  $\mathbf{k}_{1,90}$ , are derived from a single laser with wavelength  $\lambda \approx 313\,\text{nm}$ . Beams  $\mathbf{k}_{1,\text{Co}1}$  and  $\mathbf{k}_{1,\text{Co}2}$  are co-propagating such that their wave vector differences with respect to the  $\mathbf{k}_{1,90}$  beam are aligned along the trap axis. In this configuration, only the axial motional modes interact with the laser beams. The two co-propagating beams induce detuned blue and red sideband Raman transitions, respectively, when paired with the  $\mathbf{k}_{1,90}$  beam to implement the MS interaction (see Methods).

One important consideration in creating deterministic mixed-element entanglement with the MS interaction driven by multiple laser fields is the control over the relative optical phases at the ions' locations. The basis states  $|+\rangle_j, |-\rangle_j$ , and the state-dependent forces that are applied to them (see Methods) depend on the optical phases of the beams  $\mathbf{k}_{j,\text{Co}1}$ ,  $\mathbf{k}_{j,\text{Co}2}$ , and  $\mathbf{k}_{j,90}$  ( $j = 1, 2$ ) at the ion positions. Beams  $\mathbf{k}_{j,\text{Co}1}$  and  $\mathbf{k}_{j,\text{Co}2}$  are generated in the same acousto-optic modulator, one for

<sup>1</sup>National Institute of Standards and Technology, 325 Broadway, Boulder, Colorado 80305, USA. <sup>†</sup>Present addresses: JILA, University of Colorado and National Institute of Standards and Technology, and Department of Physics, University of Colorado, Boulder, Colorado 80309, USA (Y.L.); University of Washington, Department of Physics, Box 351560, Seattle, Washington 98195, USA (R.B.).





**Figure 1 | Configuration of laser beams for the mixed-element entangling gate.** For the  $^9\text{Be}^+$  ion, 313 nm laser beams (in red) simultaneously induce near-resonant red and blue sideband transitions. Similarly, for  $^{25}\text{Mg}^+$ , 280 nm beams (in green) induce sideband transitions. When all beams are applied simultaneously this implements the MS spin-spin interaction (see Methods). Each set of qubit-addressing laser beams is set up such that the wave vector differences  $\Delta k_{j,r} = k_{j,90} - k_{j,Co1}$  and  $\Delta k_{j,b} = k_{j,90} - k_{j,Co2}$  ( $j = 1, 2$ ) are aligned in the same direction along the trap axis such that only motional modes along this axis can be excited.

each ion species, and travel nearly identical paths. However, the  $k_{j,90}$  beams take a substantially different path to reach the ions' locations. Temperature drift and acoustic noise cause changes in the different beam paths that lead to phase fluctuations in the MS interaction. These fluctuations are slow on the timescale of a single gate but substantial over the course of many experiments. To suppress these effects, we embed the MS interaction in a Ramsey sequence implemented with two  $\pi/2$  carrier pulses induced by  $k_{j,Co1}$  (solid arrows) and  $k_{j,90}$  for each qubit<sup>22</sup> (Methods). The first set of pulses maps the  $|\uparrow\rangle$  and  $|\downarrow\rangle$  states of each qubit onto the  $|+\rangle_j$  and  $|-\rangle_j$  states, whose phases are synchronized with the MS interaction. The final set of pulses undoes this mapping such that the action of this sequence is independent of the path length differences as long as the differences are constant during the entire sequence. In this case, the sequence produces a phase gate  $\hat{G}$  that implements  $|\uparrow\uparrow\rangle \rightarrow |\uparrow\uparrow\rangle$ ,  $|\uparrow\downarrow\rangle \rightarrow i|\uparrow\downarrow\rangle$ ,  $|\downarrow\uparrow\rangle \rightarrow i|\downarrow\uparrow\rangle$ , and  $|\downarrow\downarrow\rangle \rightarrow |\downarrow\downarrow\rangle$ . Such a phase gate could also be implemented as in ref. 9 (on qubits

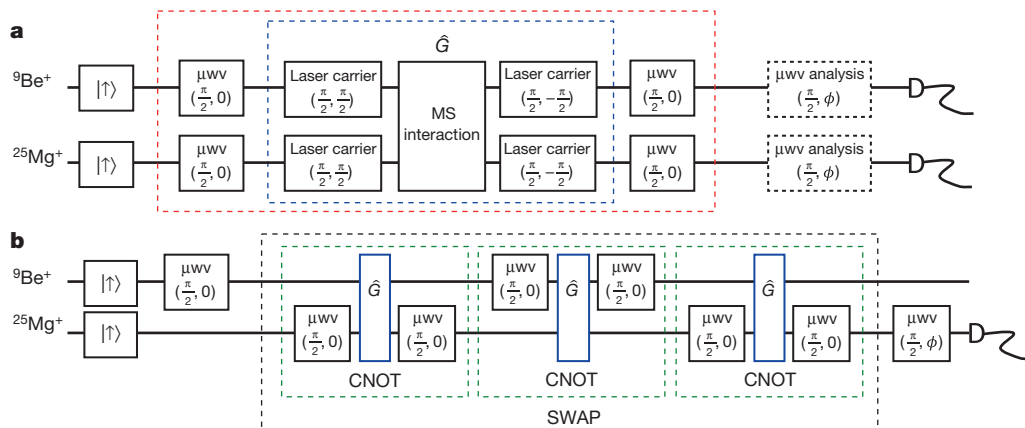
with magnetic-field-sensitive transitions). This requires fewer laser beams but adds the technical difficulty of synchronizing the state-dependent forces at the ion locations for both species.

Before applying the gate, the ions are first Doppler cooled in all three directions. The axial motional modes are further cooled to near the ground state by Raman sideband cooling on the  $^9\text{Be}^+$  ion<sup>23</sup>. State initialization into the qubits'  $|\uparrow\rangle$  states and qubit state readout are described in Methods. After each experiment repetition, we measure one of the possible states:  $|\uparrow\uparrow\rangle$ ,  $|\uparrow\downarrow\rangle$ ,  $|\downarrow\uparrow\rangle$ , or  $|\downarrow\downarrow\rangle$ .

In a first experiment, we prepare the Bell state  $\Phi_+$  with the MS interaction (Fig. 1) and determine its fidelity by measuring the qubit populations and the contrast of the parity oscillation by applying 'analysis' pulses<sup>24</sup>. The analysis pulses are laser carrier transitions induced by the non-co-propagating laser beams  $k_{j,Co1}$  and  $k_{j,90}$  such that the relative phase defining the basis states of MS interaction is stable with respect to that of the analysis pulses for each experiment repetition. We determine a Bell state fidelity of 0.979(1) (the number in parentheses is the standard error of the mean). We also create a Bell state by applying microwave carrier  $\pi/2$  pulses on each qubit before and after the operation  $\hat{G}$  (red-dashed box in Fig. 2a) achieving a fidelity of 0.964(1). Following the procedure of ref. 25, we perform a CHSH-type Bell-inequality test<sup>15</sup> on this state, achieving a sum of correlations of  $B = 2.70(2) > 2$ . This inequality, measured on an entangled system consisting of different elements, agrees with the predictions of quantum mechanics while eliminating the detection loophole but not the locality loophole<sup>25</sup>.

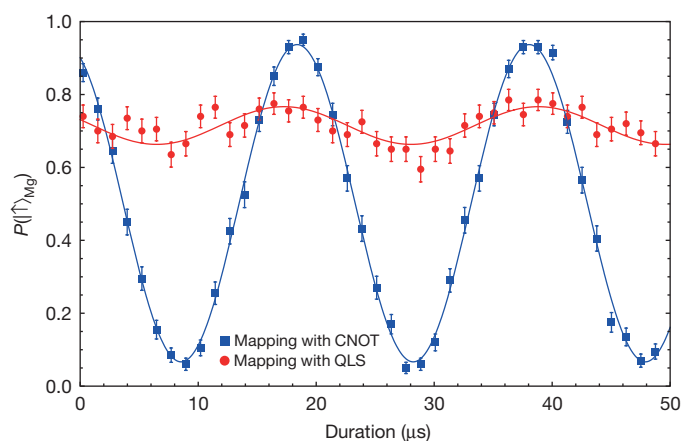
The imperfections of the entangled states can be attributed to multiple causes, which we investigate through calibration measurements and numerical simulation. We estimate the error from imperfect state preparation and detection to be  $5 \times 10^{-3}$  (see Methods). Other errors are spontaneous photon scattering<sup>26</sup> of  $^{25}\text{Mg}^+$  ( $6 \times 10^{-3}$ ) and  $^9\text{Be}^+$  ( $1 \times 10^{-3}$ ), and heating of the motional mode due to electric field noise ( $4 \times 10^{-3}$ ) (ref. 27). Other known error sources include imperfect single-qubit pulses, off-resonant coupling to spectator hyperfine states and the other motional modes, mode frequency fluctuations, qubit decoherence due to magnetic field fluctuations, laser intensity fluctuations, optical phase fluctuations, and calibration errors. Each of these sources contributes error of the order of  $10^{-3}$  or less. We find close agreement between the experimental data and numerical simulations that include the listed imperfections.

We use  $\hat{G}$  to construct a CNOT gate by applying microwave  $\pi/2$  pulses on one of the qubits before and after  $\hat{G}$  (green-dashed boxes in Fig. 2b) and use it to demonstrate qubit state mapping. The 'target' of



**Figure 2 | Pulse sequences for logic gates.** **a**, Starting with the  $|\uparrow\rangle_{\text{Be}}|\uparrow\rangle_{\text{Mg}}$  state, this pulse sequence generates a Bell state with  $\hat{G}$  (blue-dashed box) and single-qubit microwave ( $\mu\text{wv}$ ) gates. The notation  $(\theta, \phi)$  represents the rotation angle and relative phase of each gate pulse. A parity oscillation is induced by applying analysis  $\pi/2$  pulses with a variable phase  $\phi$  to the created Bell state. To demonstrate the phase insensitivity of  $\hat{G}$ , the

single-qubit gates and the analysis pulses are implemented by microwave fields that are not phase synchronized to the optical phases. **b**, Pulse sequence of a Ramsey experiment where a superposition state of a  $^9\text{Be}^+$  qubit is coherently transferred to a  $^{25}\text{Mg}^+$  qubit with a SWAP gate (black-dashed box). Given  $\hat{G}$ , either of the two qubits can be the target qubit of a CNOT gate (green-dashed boxes) by applying single-qubit  $\pi/2$  pulses to it.

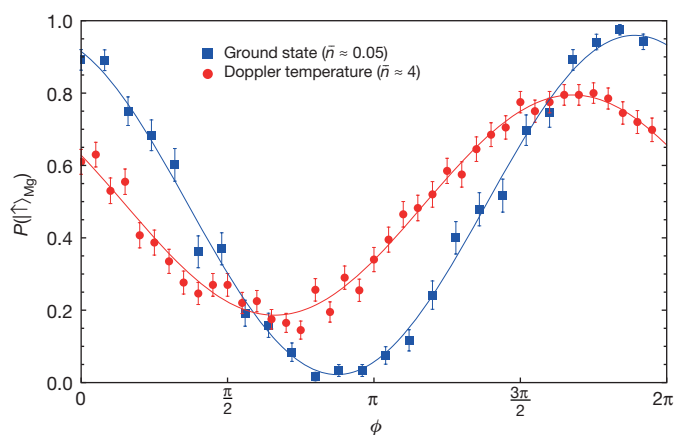


**Figure 3 | Robustness of quantum logic readout against thermal excitation.** Shown is Rabi flopping of the  ${}^9\text{Be}^+$  ion detected on the  ${}^{25}\text{Mg}^+$  ion with the motional modes cooled to Doppler temperatures using the two mapping procedures described in the text.  $P(|\uparrow\rangle_{\text{Mg}})$  is the probability of finding the  ${}^{25}\text{Mg}^+$  qubit in the  $|\uparrow\rangle$  state versus the duration of the carrier pulse on the  ${}^9\text{Be}^+$  qubit. The CNOT mapping technique, which makes use of the mixed-species gate described here, performs better than the conventional QLS procedure due to the relative insensitivity to motional excitation. Each data point represents 200 repetitions; error bars, s.e.m.

the CNOT gate is the qubit to which the single-qubit pulses are applied. The CNOT gate inherits the robustness against motional excitation from the MS gate<sup>5–8</sup>. We compare the results obtained using the CNOT gate with the method used in the conventional QLS procedure where a red-sideband  $\pi$  pulse is first applied to the  ${}^9\text{Be}^+$  ion followed by a red-sideband  $\pi$  pulse to the  ${}^{25}\text{Mg}^+$  ion<sup>14</sup>. Both procedures are calibrated for the motional mode ground state. Figure 3 shows Rabi flopping of the  ${}^9\text{Be}^+$  qubit as detected on the  ${}^{25}\text{Mg}^+$  ion, which is initially prepared in the  $|\uparrow\rangle$  state. For the ions' motional modes cooled to Doppler temperature (mean occupation number  $\bar{n} \approx 4$ ), the contrast of the conventional QLS method (red dots) is reduced compared to transfer with the CNOT gate (blue squares). In both of these mapping procedures the  ${}^9\text{Be}^+$  qubit phase information is not accessible on the  ${}^{25}\text{Mg}^+$  ion. To preserve this phase information, we construct a SWAP gate that interchanges the quantum state of the two qubits<sup>13</sup> with three CNOT gates. Figure 2b shows the pulse sequence of a Ramsey-type experiment where the first Ramsey (microwave)  $\pi/2$  pulse is applied to the  ${}^9\text{Be}^+$  ion and the second (microwave)  $\pi/2$  pulse is applied to the  ${}^{25}\text{Mg}^+$  ion after implementing the SWAP gate. Ramsey fringes for the ions' axial motional modes initialized to near the ground state ( $\bar{n} \approx 0.05$ , blue squares) and Doppler cooled ( $\bar{n} \approx 4$ , red dots) are shown in Fig. 4. The contrast at Doppler temperature is reduced because the Lamb–Dicke limit is not rigorously satisfied. Through simulation with and without the measured  ${}^{25}\text{Mg}^+$  qubit decoherence, we determine that the loss of contrast for the SWAP gate due to this decoherence is approximately 2%. For all three methods, the contrast could be somewhat improved by calibrating all gates for the given motional temperature.

We have demonstrated a mixed-element entangling gate where we employ a Ramsey sequence to suppress loss of fidelity of the output state due to low-frequency optical path length fluctuations<sup>22</sup>. Using this gate, we implement CNOT and SWAP operations between qubit elements which are relatively robust against thermal excitation of the motion. These and related techniques are potentially useful for building a large scale processor or quantum network using the advantageous properties of different ion species<sup>16,28</sup>. The entangling technique should also be applicable to qubits with optical transitions (for example,  $\text{Ca}^+$  or  $\text{Sr}^+$ ), or a combination of hyperfine qubits and optical qubits, which can also make this technique useful for readout in quantum logic clocks<sup>29</sup>.

Similar work has also been carried out at the University of Oxford<sup>30</sup> on different isotopes of  $\text{Ca}^+$  where the same laser beams



**Figure 4 | Ramsey experiments with SWAP gate.** Shown are the Ramsey fringes of the  ${}^{25}\text{Mg}^+$  qubit after initializing the  ${}^9\text{Be}^+$  ion in the  $\frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)$  state and applying the SWAP gate.  $P(|\uparrow\rangle_{\text{Mg}})$  is the probability of finding the  ${}^{25}\text{Mg}^+$  qubit in the  $|\uparrow\rangle$  state and  $\phi$  is the relative phase of the  $\pi/2$  pulse applied to the  ${}^{25}\text{Mg}^+$  qubit. The solid lines are fitted curves with contrast of 94% for the ions initialized to the ground state (mean occupation number  $\bar{n} \approx 0.05$ ) and 61% for the ions initialized to the Doppler cooling temperature ( $\bar{n} \approx 4$ ). The phase offset depends on the calibration of the SWAP gate and can be experimentally adjusted to any value. Each data point represents 200 repetitions; error bars, s.e.m.

can manipulate both isotopes simultaneously. The method presented here uses two substantially different sets of laser beams with different wavelengths, illustrating that cross-talk between operations on different species can be negligible, and could be applied to take advantage of the desirable features of each species.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 August; accepted 26 October 2015.

- Wallquist, M., Hammerer, K., Rabl, P., Lukin, M. & Zoller, P. Hybrid quantum devices and quantum engineering. *Phys. Scr.* **T137**, 014001 (2009).
- Barrett, M. D. *et al.* Sympathetic cooling of  ${}^9\text{Be}^+$  and  ${}^{24}\text{Mg}^+$  for quantum logic. *Phys. Rev. A* **68**, 042302 (2003).
- Lin, Y. *et al.* Dissipative production of a maximally entangled steady state of two quantum bits. *Nature* **504**, 415–418 (2013).
- Hume, D. B. *et al.* High-fidelity adaptive qubit detection through repetitive quantum nondemolition measurements. *Phys. Rev. Lett.* **99**, 120502 (2007).
- Sørensen, A. & Mølmer, K. Quantum computation with ions in thermal motion. *Phys. Rev. Lett.* **82**, 1971–1974 (1999).
- Sørensen, A. & Mølmer, K. Entanglement and quantum computation with ions in thermal motion. *Phys. Rev. A* **62**, 022311 (2000).
- Millburn, G. J., Schneider, S. & James, D. F. V. Ion trap quantum computing with warm ions. *Fortschr. Phys.* **48**, 801–810 (2000).
- Solano, E., de Matos Filho, R. L. & Zagury, N. Deterministic Bell states and measurement of the motional state of two trapped ions. *Phys. Rev. A* **59**, R2539–R2543 (1999).
- Leibfried, D. *et al.* Experimental demonstration of a robust, high-fidelity geometric two ion-qubit phase gate. *Nature* **422**, 412–415 (2003).
- Barenco, A. *et al.* Elementary gates for quantum computation. *Phys. Rev. A* **52**, 3457–3467 (1995).
- Bremner, M. J. *et al.* Practical scheme for quantum computation with any two-qubit entangling gate. *Phys. Rev. Lett.* **89**, 247902 (2002).
- Zhang, J., Vala, J., Sastry, S. & Whaley, K. B. Exact two-qubit universal quantum circuit. *Phys. Rev. Lett.* **91**, 027903 (2003).
- Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* 23 (Cambridge Univ. Press, 2000).
- Schmidt, P. O. *et al.* Spectroscopy using quantum logic. *Science* **309**, 749–752 (2005).
- Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
- Monroe, C. *et al.* Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Phys. Rev. A* **89**, 022317 (2014).
- Moehring, D. L. *et al.* Quantum networking with photons and trapped atoms. *J. Opt. Soc. Am. B* **24**, 300–315 (2007).
- Wineland, D. J. *et al.* Experimental issues in coherent quantum-state manipulation of trapped atomic ions. *J. Res. Natl. Inst. Stand. Technol.* **103**, 259–328 (1998).

19. Schulte, M., Lörch, N., Leroux, I. D., Schmidt, P. O. & Hammerer, K. Quantum algorithmic readout in multi-ion clocks. Preprint at <http://arXiv.org/abs/1501.06453> (2015).
20. Blakestad, R. B. *et al.* Near-ground-state transport of trapped-ion qubits through a multidimensional array. *Phys. Rev. A* **84**, 033314 (2011).
21. Langer, C. *et al.* Long-lived qubit memory using atomic ions. *Phys. Rev. Lett.* **95**, 060502 (2005).
22. Lee, P. J. *et al.* Phase control of trapped ion quantum gates. *J. Opt. B* **7**, S371–S383 (2005).
23. Monroe, C. *et al.* Resolved-sideband Raman cooling of a bound atom to the 3D zero-point energy. *Phys. Rev. Lett.* **75**, 4011–4014 (1995).
24. Sackett, C. A. *et al.* Experimental entanglement of four particles. *Nature* **404**, 256–259 (2000).
25. Rowe, M. A. *et al.* Experimental violation of a Bell's inequality with efficient detection. *Nature* **409**, 791–794 (2001).
26. Ozeri, R. *et al.* Errors in trapped-ion quantum gates due to spontaneous photon scattering. *Phys. Rev. A* **75**, 042329 (2007).
27. Turchette, Q. A. *et al.* Heating of trapped ions from the quantum ground state. *Phys. Rev. A* **61**, 063418 (2000).
28. Monroe, C. & Kim, J. Scaling the ion trap quantum processor. *Science* **339**, 1164–1169 (2013).
29. Chou, C. W., Hume, D. B., Koelemeij, J. C. J., Wineland, D. J. & Rosenband, T. Frequency comparison of two high-accuracy  $\text{Al}^+$  optical clocks. *Phys. Rev. Lett.* **104**, 070802 (2010).
30. Ballance, C. J. *et al.* Hybrid quantum logic and a test of Bell's inequality using two different atomic isotopes. *Nature* <http://dx.doi.org/10.1038/nature16184> (this issue).

**Acknowledgements** We thank J. Bollinger and D. Hume for comments on the manuscript. This work was supported by the Office of the Director of National Intelligence (ODNI) Intelligence Advanced Research Projects Activity (IARPA), ONR, and the NIST Quantum Information Program. Y.W. was supported by the US Army Research Office through MURI grant W911NF-11-1-0400. This paper is a contribution by NIST and not subject to US copyright.

**Author Contributions** T.R.T. and J.P.G. conceived and designed the experiments, developed components of the experimental apparatus, and collected and analysed data. T.R.T. wrote the manuscript. Y.L., Y.W., and R.B. contributed to the development of experimental apparatus. D.L. and D.J.W. directed the experiment. All authors provided important suggestions for the experiments, discussed the results, and contributed to the editing of the manuscript.

**Additional Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.R.T. ([tingrei.tan@nist.gov](mailto:tingrei.tan@nist.gov)).



## METHODS

**Geometric phase gate.** The Mølmer-Sørensen (MS) protocol<sup>5–8</sup> requires simultaneous excitation of a blue sideband transition with a detuning of  $\delta$  and a red sideband transition with a detuning of  $-\delta$  for a selected motional mode (Fig. 1). The excitation creates a forced harmonic oscillator interaction that displaces the motional wavefunction in phase space in a manner that is dependent on the internal qubit states. If the different displacements enclose a loop, the qubit states pick up a geometric phase proportional to the state-dependent area of the enclosed loop. We create an entangling logic gate by choosing appropriate geometric phase differences between different qubit states.

Laser fields are used to induce coherent stimulated-Raman transitions between the qubit states of each ion and the shared quantized degrees of motion<sup>18</sup>. For each qubit we can excite carrier transitions  $|j, n\rangle \leftrightarrow |j, n\rangle$  that induce spin-flips without changing the motional Fock state  $n$ . A blue (red) sideband excitation flips the spin while adding (removing) a quantum of motion by detuning the fields from the carrier transition frequency by the motional frequency. The relative frequencies, phases, and intensities of each set of laser beams (Fig. 1) for each qubit can be adjusted with acousto-optic modulators (AOMs), which are computer controlled. The  $^9\text{Be}^+$  Raman laser beams with a wavelength of  $\lambda \approx 313$  nm are approximately 480 GHz red detuned from the  $S_{1/2}$  to  $P_{1/2}$  electronic state transition. The  $\lambda \approx 280$  nm Raman laser beams for  $^{25}\text{Mg}^+$  ion are approximately 160 GHz blue detuned from the  $S_{1/2}$  to  $P_{3/2}$  electronic state transition. Carrier transitions can also be implemented by microwave fields delivered from an antenna located outside the vacuum chamber.

After transforming into the respective interaction frames of both qubits as well as that of the shared motional mode of motion, and dropping high-frequency terms in the rotating-wave approximation, we can write the interaction in the Lamb-Dicke limit as<sup>18</sup>

$$H = \hbar \sum_{j=1,2} \Omega_j \hat{\sigma}_j^{\dagger} \left( \hat{a} e^{-i(\delta t - \phi_{j,r})} + \hat{a}^{\dagger} e^{i(\delta t + \phi_{j,b})} \right) + h.c.$$

where  $j = 1, 2$  denotes the two different ion species,  $\Omega_j = \eta_j \Omega_{0,j}$  where  $\Omega_{0,j}$  is the carrier resonance Rabi frequency. The Lamb-Dicke parameter  $\eta_j$  is equal to  $\Delta k_j z_{0,j} b_j$ , where  $b_j$  is the mode amplitude of the  $j$ th ion and  $z_{0,j} = \sqrt{\hbar/2m_j\omega_z}$ ,  $m_j$  is the mass and  $\omega_z$  is the frequency of the selected normal mode. The spin raising operator is  $\hat{\sigma}_j^{\dagger}$  and  $\hat{a}^{\dagger}$  is the creation operator for the relevant (harmonic) motional mode.

The phases of the red (r) and blue (b) sideband interactions are  $\phi_{j,r(b)} = \Delta k_{j,r(b)} X_{0,j} + \Delta \phi_{j,r(b)}$  where  $\Delta k_{j,r(b)}$  and  $\Delta \phi_{j,r(b)}$  are the differences in wave vectors and phases of the optical fields driving the red and blue sideband transitions, respectively, and  $X_{0,j}$  is the equilibrium position for the  $j$ th ion. After setting  $\Omega_1 = \Omega_2 = \Omega$  and  $\delta_1 = \delta_2 = \delta$ , and writing  $\phi_{M,j} = (\phi_{j,r} - \phi_{j,b})/2$ , the geometric phases accumulated after a duration of  $t_{\text{MS}} = 2\pi/\delta$  for the four  $|+\rangle_j$  and  $|-\rangle_j$  basis states (defined as the eigenstates of  $\hat{\sigma}_{\phi,j} = \cos((\phi_{j,r} + \phi_{j,b})/2)\hat{\sigma}_{x,j} - \sin((\phi_{j,r} + \phi_{j,b})/2)\hat{\sigma}_{y,j}$ ) are

$$\begin{aligned} \varphi_{|+,+\rangle,|-, -\rangle} &= \frac{8\pi\Omega^2}{\delta^2} \cos^2\left(\frac{\phi_{M,1} - \phi_{M,2}}{2}\right) \\ \varphi_{|+,-\rangle,|-, +\rangle} &= \frac{8\pi\Omega^2}{\delta^2} \sin^2\left(\frac{\phi_{M,1} - \phi_{M,2}}{2}\right) \end{aligned} \quad (1)$$

To maximize entangling gate speed, the geometric phases for the different parity qubit states in equation (1) are set to differ by  $\pi/2$ . This is accomplished by adjusting the phases of the radio frequencies driving the AOMs.

There are two axial modes: the lower-frequency mode ( $\omega_z = 2\pi \times 2.5$  MHz), where the ions oscillate in phase, and the higher-frequency mode ( $2\pi \times 5.4$  MHz), where the ions oscillate out of phase. The Lamb-Dicke parameters for the  $^9\text{Be}^+$

( $^{25}\text{Mg}^+$ ) ion are 0.156 (0.265) and 0.269 (0.072), respectively, for the two modes. We use the in-phase mode for our demonstration because the  $^{25}\text{Mg}^+$  ion has a larger normal mode amplitude compared to the out-of-phase mode. This results in less spontaneous emission error for a given strength of the state-dependent force. Gate time  $t_{\text{MS}}$  is approximately 35  $\mu\text{s}$ .

**Calibration procedure for phase gate  $\hat{G}$ .** To produce the phase gate  $\hat{G}$ , the phases of the  $\pi/2$  pulses for the Ramsey sequence must be referenced to the basis states of the MS interaction defined by the optical phases. The phases must also account for the AC Stark shifts induced by the laser beams that are used for the MS interaction.

To calibrate these phases, we first perform the pulse sequence shown in the blue-dashed box of Fig. 2a with the MS interaction pulses detuned far off-resonant from the red and blue sideband transitions such that they only induce AC Stark shifts on the qubits. Starting with the input state  $|\uparrow\uparrow\rangle$ , we set the phases of the final  $\pi/2$  laser pulses such that the action of this pulse sequence returns each qubit to the  $|\uparrow\uparrow\rangle$  state. Then, we perform this sequence with the MS interactions correctly tuned and vary the phases of the MS interactions. Again, in this case we look for the phase that maps the input state  $|\uparrow\uparrow\rangle$  back to itself. We verify the action of this  $\hat{G}$  operation by creating a Bell state with the pulse sequence shown in Fig. 2a.

**Qubit state preparation and readout.** For qubit state preparation, the  $^9\text{Be}^+$  ion is optically pumped to the  $|2, 2\rangle$  state followed by Doppler cooling implemented by driving the  $S_{1/2}|2, 2\rangle \leftrightarrow P_{3/2}|3, 3\rangle$  cycling transition with  $\sigma^+$  polarized light. Similarly, we optically pump the  $^{25}\text{Mg}^+$  ion to the  $|3, 3\rangle$  state and apply Doppler cooling on the  $S_{1/2}|3, 3\rangle \leftrightarrow P_{3/2}|4, 4\rangle$  transition. For ground state initialization of the axial motional modes, Raman sideband cooling is applied to the  $^9\text{Be}^+$  ion<sup>23</sup>. To transfer the  $^9\text{Be}^+$   $|2, 2\rangle$  state to the  $|1, 1\rangle = |\uparrow\rangle_{\text{Be}}$  state, we use microwave composite pulse sequences that are robust against transition detuning errors. These consist of resonant  $(\frac{\pi}{2}, 0), (\frac{3\pi}{2}, \frac{\pi}{2}), (\frac{\pi}{2}, 0)$  pulses<sup>31</sup>, where the first entry denotes the angle the state is rotated about a vector in the  $x$ - $y$  plane of the Bloch sphere and the second angle represents the azimuthal angle for the rotation axis. With analogous sequences, we first transfer the  $^{25}\text{Mg}^+$  from the  $|3, 3\rangle$  state to the  $|2, 2\rangle$  state, and then to the  $|3, 1\rangle = |\uparrow\rangle_{\text{Mg}}$  state.

State-dependent resonance-fluorescence detection is accomplished with an achromatic lens system designed for 313 nm and 280 nm (ref. 32). We sequentially image each ion's fluorescence onto a photomultiplier tube. After reversing the initial mapping procedures to put the  $|\uparrow\rangle$  states back in the respective cycling transition ground states, we apply the Doppler cooling beams. The fluorescing or 'bright' state of this protocol therefore corresponds to the  $|\uparrow\rangle$  state of each ion. The  $|\downarrow\rangle$  state of each qubit is transferred to  $|1, -1\rangle$  and  $|2, -2\rangle$  for the  $^9\text{Be}^+$  and  $^{25}\text{Mg}^+$ , respectively, with microwave carrier  $\pi$  pulses. These states are 'dark' to the detection beams and correspond to the  $|\downarrow\rangle$  state. This 'shelving' technique is used to minimize the overlap of the bright and dark state photon count probability distributions. With detection durations of 330  $\mu\text{s}$  for  $^9\text{Be}^+$  and 200  $\mu\text{s}$  for  $^{25}\text{Mg}^+$ , we detect on average 30 photons for each ion when they are in the bright state and 3.5 photons (predominantly from background light) when they are in the dark state. The qubit state is determined by choosing a photon count threshold such that the states are maximally distinguished. The state preparation and detection error of  $5 \times 10^{-3}$  reported in the main text includes errors due to the threshold detection protocol (false determination of each detected state being in the other state) and the infidelities of the microwave transfer pulses.

**Sample size.** No statistical methods were used to predetermine sample size.

- Levitt, M. H. Composite pulses. *Prog. Nucl. Magn. Reson. Spectrosc.* **18**, 61–122 (1986).
- Huang, P. & Leibfried, D. Achromatic catadioptric microscope objective in deep ultraviolet with long working distance. *Proc. SPIE* **5524**, 125–133 (2004).

# Hybrid quantum logic and a test of Bell's inequality using two different atomic isotopes

C. J. Ballance<sup>1</sup>, V. M. Schäfer<sup>1</sup>, J. P. Home<sup>1</sup>, D. J. Szwer<sup>1</sup>, S. C. Webster<sup>1</sup>, D. T. C. Allcock<sup>1</sup>, N. M. Linke<sup>1</sup>, T. P. Harty<sup>1</sup>, D. P. L. Aude Craik<sup>1</sup>, D. N. Stacey<sup>1</sup>, A. M. Steane<sup>1</sup> & D. M. Lucas<sup>1</sup>

Entanglement is one of the most fundamental properties of quantum mechanics<sup>1–3</sup>, and is the key resource for quantum information processing<sup>4,5</sup> (QIP). Bipartite entangled states of identical particles have been generated and studied in several experiments, and post-selected or heralded entangled states involving pairs of photons, single photons and single atoms, or different nuclei in the solid state, have also been produced<sup>6–12</sup>. Here we use a deterministic quantum logic gate to generate a ‘hybrid’ entangled state of two trapped-ion qubits held in different isotopes of calcium, perform full tomography of the state produced, and make a test of Bell's inequality with non-identical atoms. We use a laser-driven two-qubit gate<sup>13</sup>, whose mechanism is insensitive to the qubits' energy splittings, to produce a maximally entangled state of one <sup>40</sup>Ca<sup>+</sup> qubit and one <sup>43</sup>Ca<sup>+</sup> qubit, held 3.5 micrometres apart in the same ion trap, with  $99.8 \pm 0.6$  per cent fidelity. We test the CHSH (Clauser–Horne–Shimony–Holt)<sup>14</sup> version of Bell's inequality for this novel entangled state and find that it is violated by 15 standard deviations; in this test, we close the detection loophole<sup>8</sup> but not the locality loophole<sup>7</sup>. Mixed-species quantum logic is a powerful technique for the construction of a quantum computer based on trapped ions, as it allows protection of memory qubits while other qubits undergo logic operations or are used as photonic interfaces to other processing units<sup>15,16</sup>. The entangling gate mechanism used here can also be applied to qubits stored in different atomic elements; this would allow both memory and logic gate errors caused by photon scattering to be reduced below the levels required for fault-tolerant quantum error correction, which is an essential prerequisite for general-purpose quantum computing.

For Schrödinger, entanglement was “the characteristic trait of quantum mechanics”<sup>1</sup> and it has been at the heart of debates about the foundations of quantum mechanics since the framing of the Einstein–Podolsky–Rosen paradox<sup>2</sup>. The theoretical work of Bell<sup>3</sup>, and of Clauser *et al.*<sup>14</sup>, established an experimental test which could be used to rule out local hidden-variable theories on the basis of correlations between measured properties of entangled particles, and numerous experiments, starting with that of Freedman and Clauser, have confirmed the predictions of quantum mechanics<sup>6–10</sup>. Tests of Bell's inequality with trapped ions were the first to close the so-called ‘detection loophole’; hitherto these trapped-ion tests had been exclusively carried out with identical atoms<sup>8,17,18</sup>. The entanglement explored in tests of Bell's inequality is typically an entanglement between distinguishable particles, in the strict quantum mechanical sense, but when the particles are identical in their internal structure and state, they are distinguishable only through their spatial localization. By employing different isotopes, our experiments involve entities that are also distinguishable by many internal properties, such as baryon number, mass, spin and resonant frequencies.

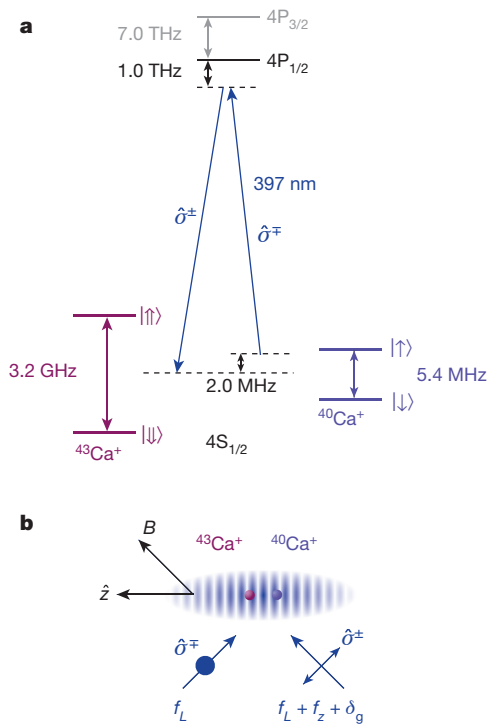
Apart from its intrinsic interest, entanglement is a central resource for quantum information applications, such as quantum cryptography<sup>5</sup> and quantum computing<sup>4</sup>. Trapped atomic ions are one of the most

promising technologies for the implementation of quantum computation; several demonstrations of simple multi-qubit algorithms have been made<sup>19</sup> and the elementary set of quantum logic operations has recently been demonstrated with the precision required for the implementation of fault-tolerant techniques<sup>20,21</sup>. Scaling up trapped-ion systems to the large numbers of qubits required for useful QIP and quantum simulation will almost certainly require the use of more than one species of ion, both for the purpose of sympathetic laser-cooling (which allows independent control of the external and internal atomic degrees of freedom)<sup>15,22,23</sup> and for providing robust memory qubits. The best memory qubits reside in hyperfine ground states<sup>20,24</sup>, which have essentially infinite lifetimes against spontaneous decay, but are vulnerable to the scattering of a single photon of resonant laser light. In a complex, multi-zone, ion trap processor it will be difficult to shield the memory qubits sufficiently well from resonant laser beams, hence it will be useful to employ different species of ion—for example, as memory and logic qubits—and a high-fidelity entangling gate operation between the two species will be invaluable. A significant initial step was the demonstration of coherent state transfer between different species in the context of precision metrology<sup>25,26</sup>. The relative merits of using different isotopes versus different elements are discussed below.

In the present work, we entangle qubits stored in two different isotopes of calcium. The <sup>40</sup>Ca<sup>+</sup> qubit is stored in the Zeeman-split ground level,  $(|\downarrow\rangle, |\uparrow\rangle) = (4S_{1/2}^{-1/2}, 4S_{1/2}^{+1/2})$ , and the <sup>43</sup>Ca<sup>+</sup> qubit is stored in the hyperfine ground states  $(|\downarrow\rangle, |\uparrow\rangle) = (4S_{1/2}^{4,+4}, 4S_{1/2}^{3,+3})$ , see Fig. 1. The qubit energy splittings differ by some three orders of magnitude ( $f_{\uparrow} \approx 5.4$  MHz,  $f_{\downarrow} \approx 3.2$  GHz), but they may nevertheless be efficiently coupled via the two-qubit gate mechanism of ref. 13, in which the ‘travelling standing wave’ from a pair of far-detuned laser beams exerts a qubit-state-dependent force on the ions whose magnitude  $F$  is largely independent of the qubit frequency. The force originates from a spatially varying light shift, oscillates at the difference frequency  $\delta$  between the two beams and, when  $\delta = f_z + \delta_g$  is set close to the resonant frequency  $f_z$  of a normal mode of motion of the two-ion crystal, a two-qubit phase gate may be implemented by applying the force for a time of  $1/\delta_g$ . An advantage of this type of gate is that the phase of the optical field does not need to be referenced to either of the qubit phases (see Methods); this makes scaling the system easier because the relative optical phase does not need to be controlled between different trap zones, or during time delays between gates.

An important difference in the gate mechanism compared with the case of identical ions is that the forces on corresponding qubit states differ ( $F_{\uparrow} \neq F_{\downarrow}$  and  $F_{\downarrow} \neq F_{\uparrow}$ ) so that, in general, the four possible qubit states ( $\uparrow\uparrow, \uparrow\downarrow, \downarrow\uparrow, \downarrow\downarrow$ ) each acquire different phases. We choose to implement the gate operation in two halves, each of duration  $t_g/2 = 1/\delta_g$ , separated by spin-flip operations ( $\pi$  pulses) on the qubits (Fig. 2a). This symmetrizes the gate operation  $\hat{G}$  such that the relative phases acquired by the four states are  $(0, \Phi, \Phi, 0)$ . By setting the laser power (that is, the

<sup>1</sup>Department of Physics, University of Oxford, Clarendon Laboratory, Parks Road, Oxford OX1 3PU, UK.

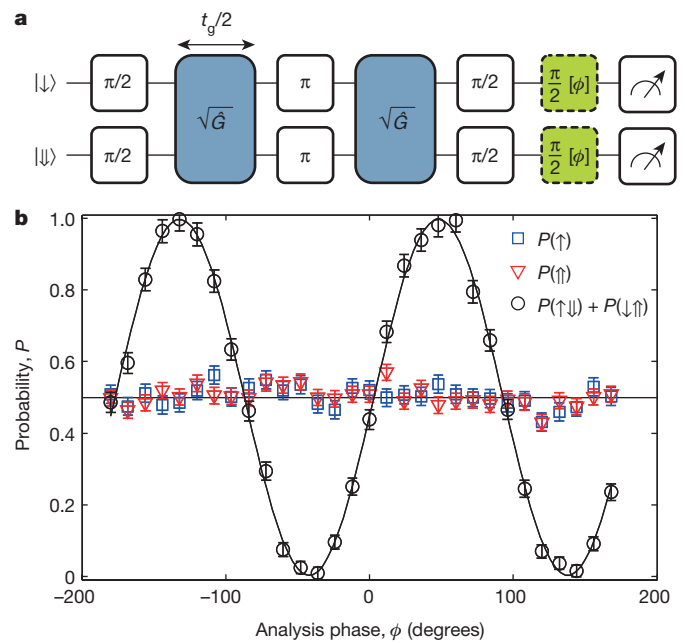


**Figure 1 | Calcium ion energy levels and experimental geometry.** **a**, Qubit states and Raman transitions in  $^{43}\text{Ca}^+$  (purple, left) and  $^{40}\text{Ca}^+$  (violet, right). The two Raman beams ( $\hat{\sigma}^{\pm}$  and  $\hat{\sigma}^{\mp}$ , blue, centre) have a frequency of  $\sim f_L$ , a mean detuning of  $\Delta = -1.04$  THz from the  $4S_{1/2} \leftrightarrow 4P_{1/2}$  (397 nm) transition, and a difference frequency of  $\delta = f_z + \delta_g \approx 2.0$  MHz. **b**, Raman gate beam geometry. The two perpendicular beams are aligned to set the lattice  $\mathbf{k}$  vector parallel to the trap axis  $\hat{z}$ . The beams have waist radii  $w = 27 \mu\text{m}$ , a power of  $\sim 5$  mW each, and orthogonal linear polarizations as indicated. A third,  $\pi$ -polarized, Raman beam (not shown) co-propagates with the  $\sigma^+$  beam and is used for sub-Doppler sideband cooling and single-qubit operations on  $^{40}\text{Ca}^+$ . The quantization axis is set by a magnetic field  $B \approx 0.2$  mT. The diagram is not to scale: the ions are separated by  $3.5 \mu\text{m}$ , which is 12.5 periods of the standing wave, and around 20,000 times the atomic radius of calcium.

effective Rabi frequency) and gate detuning  $\delta_g$  appropriately, such that  $\Phi = \pi/2$ , and enclosing the gate operation in a Ramsey interferometer (two pairs of  $\pi/2$  pulses), we can generate the maximally entangled Bell state  $(| \downarrow \downarrow \rangle + | \uparrow \uparrow \rangle)/\sqrt{2}$  from the initial state  $| \downarrow \downarrow \rangle$ . The  $\pi$  pulses also protect the qubits against dephasing due to slow ( $\gg t_g$ ) variations in magnetic fields.

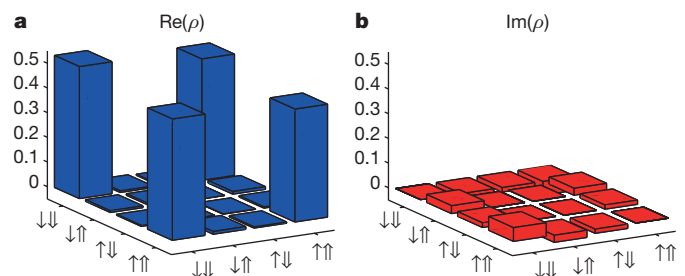
In our experiment, we implement the gate using the in-phase axial motional mode (at  $f_z = 2.00$  MHz) of a linear Paul trap<sup>27</sup>, with the ion separation ( $3.5 \mu\text{m}$ ) equal to a half-integer number of standing wavelengths, thus exciting the motion maximally for the  $| \uparrow \downarrow \rangle$  and  $| \downarrow \uparrow \rangle$  states. The Lamb–Dicke parameters for the two different isotopes are  $\eta_{40} = 0.121$  and  $\eta_{43} = 0.126$ . After initial Doppler cooling, both axial modes are cooled close to their ground states (mean occupation number  $\bar{n} < 0.1$ ) by Raman sideband cooling applied to the  $^{40}\text{Ca}^+$  ion, which sympathetically cools the  $^{43}\text{Ca}^+$  ion<sup>28</sup>. Both qubits are initialized by optical pumping, after which we apply the gate sequence shown in Fig. 2a, using a gate duration  $t_g = 27.4 \mu\text{s}$ . Single-qubit  $\pi/2$  and  $\pi$  pulses, for the spin-echo and tomography operations, are applied using co-propagating Raman beams (for  $^{40}\text{Ca}^+$ ) and microwaves (for  $^{43}\text{Ca}^+$ ). The ordering of the ion pair in the trap was kept constant over the time taken to acquire the full data set, to guard against systematic effects associated with ion position (see Methods). We implement individual single-shot qubit readout by state-selectively shelving both ions to the  $3D_{5/2}$  level simultaneously, then detecting the ions' fluorescence sequentially in two photomultiplier counting periods (see Methods).

From the contrast of the parity fringes shown in Fig. 2b, and a measurement of the qubit populations before the analysis pulses<sup>13</sup>,



**Figure 2 | Entangling gate sequence and results.** **a**, Gate sequence, showing the operations applied to the  $^{40}\text{Ca}^+$  (upper line) and  $^{43}\text{Ca}^+$  (lower line) qubits, where  $\hat{G}$  is the gate operation. The final state analysis (tomography)  $\pi/2$  pulses shown in green are optional; by scanning their phase  $\phi$  we can diagnose the state produced by the gate. **b**, Qubit populations and parity signal after correcting for readout errors (see Methods). The individual qubit populations (open squares and inverted open triangles) are consistent with  $1/2$ , as expected for the Bell state  $(| \downarrow \downarrow \rangle + | \uparrow \uparrow \rangle)/\sqrt{2}$ . The parity signal  $P(\uparrow\downarrow) + P(\downarrow\uparrow)$  (open circles), that is, the probability of the two qubits being in opposite states, should oscillate between 0 and 1 as  $\sin(2\phi)$  for a perfect Bell state. From the contrast of the parity signal and a measurement of the populations without the analysis pulses, we infer a Bell state fidelity of 99.8(6)%. The error bars show  $1\sigma$  statistical errors.

we estimate the fidelity of the Bell state produced by the gate to be  $\mathcal{F} = 99.8(6)\%$ , where the error (0.6%) is dominated by statistical uncertainty. Known contributions to the gate error are significantly smaller<sup>27</sup> than the statistical uncertainty; for example, the photon scattering error at the  $\Delta = -1.04$  THz Raman detuning used is estimated to be approximately 0.1%. Since the two qubits may be rotated independently by addressing them in frequency space, we can also perform full tomography of the entangled state and extract the density matrix (Fig. 3); the density matrix is consistent with that for the desired Bell state, to within the systematic errors from the imperfect tomography pulses, and gives a separate estimate of the fidelity  $\mathcal{F} = 99(1)\%$ . In both cases,  $\mathcal{F}$  represents the fidelity of the entangling gate operation; it excludes errors due to



**Figure 3 | Density matrix of the mixed-isotope Bell state.** **a**, **b**, The plots show the real (**a**) and imaginary (**b**) parts of the density matrix  $\rho$ , after correcting for qubit readout errors (see Methods). These were measured by rotating each qubit independently to perform full quantum state tomography. We used a maximum likelihood method to find the density matrix that best represents the experimental data. This gives a separate estimate of the gate fidelity, 99(1)%.



**Table 1 | Bell/CHSH inequality test results, using the mixed-isotope entangled state**

$\theta_a (^{40}\text{Ca}^+)$	$\pi/4$	$3\pi/4$	$\pi/4$	$3\pi/4$
$\theta_b (^{43}\text{Ca}^+)$	$\pi/2$	$\pi/2$	0	0
$E(\theta_a, \theta_b)$	0.565(7)	0.530(7)	0.560(7)	−0.573(8)

The qubits a and b are independently rotated through angles  $(\theta_a, \theta_b) = (\pi/4, 3\pi/4)$  and  $(\theta_b, \theta_a) = (\pi/2, 0)$ , and for each combination of angles the correlation function  $E(\theta_a, \theta_b)$  is measured, with results shown. ( $E$  is defined as in ref. 17.) The CHSH parameter is given by  $S = |E(\theta_a, \theta_b) + E(\theta_a, \theta_b') + E(\theta_a', \theta_b) - E(\theta_a', \theta_b')| = 2.228(15) > 2$ , thus violating Bell's inequality for this system of non-identical atoms. The state detection errors are sufficiently small (approximately 6%, see Methods) that it is not necessary to make a fair-sampling assumption. For each angle setting, 4,000 measurements were made.

state preparation and readout, which we characterize in independent experiments (see Methods).

To perform a test of the CHSH version of Bell's inequality, we follow the gate sequence with further independent single-qubit rotations and measurements. The single-qubit rotations have constant phase  $\phi$  but varying rotation angle  $\theta$ . From these measurements we determine the two-particle correlation functions with results shown in Table 1. As is well known, the maximal CHSH parameter  $S$  allowed by local hidden-variable theories is 2, whereas quantum mechanics allows  $S \leq 2\sqrt{2}$ . In order to avoid having to make a fair-sampling assumption, we do not correct for qubit readout errors in these experiments. The finite detection error then limits the CHSH parameter to a detectable maximum  $S_{\text{max}} = 2.236(7)$  for a perfect Bell state; our results give  $S = 2.228(15)$ , consistent with  $S_{\text{max}}$  to within the stated uncertainties, and violating the CHSH inequality by approximately 15%.

The mixed-species quantum logic gate that we have demonstrated has allowed us to create a novel entangled state, leading to the first test of a Bell inequality violation between isolated non-identical atoms. As an application, the two isotopes used here could be employed for scalable quantum computing architectures based on trapped ions; hyperfine qubits in  $^{43}\text{Ca}^+$  at present constitute the best single-qubit memories (coherence time  $T_2^* \approx 1$  min)<sup>20</sup>, whereas the simpler atomic structure of  $^{40}\text{Ca}^+$  is well suited for use as a 'photonic interconnect' qubit<sup>16</sup>. There are technical advantages to using ions of similar mass for sympathetic cooling and ion transport in multi-zone traps. However, while the relatively small isotope shifts ( $\sim 1$  GHz) allow the convenient use of the same laser systems for manipulation of both species, they may provide insufficient protection of qubits from stray resonant light unless tightly focused beams are used<sup>18,28</sup>. Therefore in the long term it may be necessary to use different atomic elements<sup>22</sup>. The gate mechanism employed here is independent of the qubit frequency and thus can also be used to couple qubits stored in different elements, provided that the Raman laser fields exert sufficient force on both qubits. We note that  $\text{Ca}^+$  and  $\text{Sr}^+$  ions are an attractive choice in this respect: the  $4S_{1/2} \leftrightarrow 4P_{1/2}$  transition in  $\text{Ca}^+$  is separated from the  $4S_{1/2} \leftrightarrow 4P_{3/2}$  transition in  $\text{Sr}^+$  by 20 THz. A Raman laser detuning of  $\Delta = -8$  THz (comparable to that used in our recent  $^{43}\text{Ca}^+ - ^{43}\text{Ca}^+$  two-qubit gate experiments<sup>21</sup>) would enable the implementation of a mixed-species logic gate with a photon-scattering error of  $\sim 10^{-4}$ , substantially below the error threshold for fault-tolerant operations<sup>29</sup>.

Similar experiments using trapped-ion qubits stored in two different elements ( $^9\text{Be}^+$  and  $^{25}\text{Mg}^+$ ) have recently been carried out in the NIST Ion Storage Group<sup>30</sup>. We note that after the submission of the present manuscript, a CHSH-Bell test that closes both detection and locality loopholes, using heralded entanglement of remote electron spins, was reported<sup>31</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 17 August; accepted 26 October 2015.**

1. Schrödinger, E. Discussion of probability relations between separated systems. *Math. Proc. Camb. Phil. Soc.* **31**, 555–563 (1935).

2. Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
3. Bell, J. S. On the Einstein-Podolsky-Rosen paradox. *Physics* **1**, 195–200 (1964).
4. Deutsch, D. Quantum theory, the Church-Turing principle and the universal quantum computer. *Proc. R. Soc. Lond. A* **400**, 97–117 (1985).
5. Ekert, A. K. Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.* **67**, 661–663 (1991).
6. Freedman, S. J. & Clauser, J. F. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* **28**, 938–941 (1972).
7. Aspect, A., Grangier, P. & Roger, G. Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: a new violation of Bell's inequalities. *Phys. Rev. Lett.* **49**, 91–94 (1982).
8. Rowe, M. A. et al. Experimental violation of a Bell's inequality with efficient detection. *Nature* **409**, 791–794 (2001).
9. Moehring, D. L., Madsen, M., Blinov, B. & Monroe, C. Experimental Bell inequality violation with an atom and a photon. *Phys. Rev. Lett.* **93**, 090410 (2004).
10. Giustina, M. et al. Bell violation using entangled photons without the fair-sampling assumption. *Nature* **497**, 227–230 (2013).
11. Christensen, B. G. et al. Detection-loophole-free test of quantum nonlocality, and applications. *Phys. Rev. Lett.* **111**, 130406 (2013).
12. Pfaff, W. et al. Demonstration of entanglement-by-measurement of solid-state qubits. *Nature Phys.* **9**, 29–33 (2013).
13. Leibfried, D. et al. Experimental demonstration of a robust, high-fidelity geometric two ion-qubit phase gate. *Nature* **422**, 412–415 (2003).
14. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
15. Wineland, D. J. et al. Experimental issues in coherent quantum-state manipulation of trapped atomic ions. *J. Res. Natl Inst. Stand. Technol.* **103**, 259–328 (1998).
16. Monroe, C. & Kim, J. Scaling the ion trap quantum processor. *Science* **339**, 1164–1169 (2013).
17. Matsukevich, D. N., Maunz, P., Moehring, D. L., Olmschenk, S. & Monroe, C. Bell inequality violation with two remote atomic qubits. *Phys. Rev. Lett.* **100**, 150404 (2008).
18. Lanyon, B. P. et al. Experimental violation of multipartite Bell inequalities with trapped ions. *Phys. Rev. Lett.* **112**, 100403 (2014).
19. Blatt, R. & Wineland, D. J. Entangled states of trapped atomic ions. *Nature* **453**, 1008–1015 (2008).
20. Harty, T. P. et al. High-fidelity preparation, gates, memory, and readout of a trapped-ion quantum bit. *Phys. Rev. Lett.* **113**, 220501 (2014).
21. Ballance, C. J., Harty, T. P., Linke, N. M. & Lucas, D. M. High-fidelity two-qubit quantum logic gates using trapped calcium-43 ions. Preprint at <http://arxiv.org/abs/1406.5473> (2014).
22. Barrett, M. D. et al. Sympathetic cooling of  $^9\text{Be}^+$  and  $^{24}\text{Mg}^+$  for quantum logic. *Phys. Rev. A* **68**, 042302 (2003).
23. Home, J. P. et al. Complete methods set for scalable ion trap quantum information processing. *Science* **325**, 1227–1230 (2009).
24. Langer, C. et al. Long-lived qubit memory using atomic ions. *Phys. Rev. Lett.* **95**, 060502 (2005).
25. Schmidt, P. O. et al. Spectroscopy using quantum logic. *Science* **309**, 749–752 (2005).
26. Hume, D. B., Rosenband, T. & Wineland, D. J. High-fidelity adaptive qubit detection through repetitive quantum nondemolition measurements. *Phys. Rev. Lett.* **99**, 120502 (2007).
27. Ballance, C. J. *High-Fidelity Quantum Logic in  $\text{Ca}^+$* . D.Phil. thesis, Univ. Oxford (2014).
28. Home, J. P. et al. Memory coherence of a sympathetically cooled trapped-ion qubit. *Phys. Rev. A* **79**, 050305 (2009).
29. Fowler, A. G., Mariantoni, M., Martinis, J. M. & Cleland, A. N. Surface codes: towards practical large-scale quantum computation. *Phys. Rev. A* **86**, 032324 (2012).
30. Tan, T. R. et al. Multi-element logic gates for trapped-ion qubits. *Nature* <http://dx.doi.org/10.1038/nature16186> (this issue).
31. Hensen, B. et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).

**Acknowledgements** This work was supported by the UK EPSRC 'Networked Quantum Information Technology' Hub and the US Army Research Office (contract W911NF-14-1-0217). D.M.L. thanks A. Castillo and E. A. Castillo for their hospitality while revising the manuscript.

**Author Contributions** All authors contributed to the development of the apparatus and/or the design of the experiments. J.P.H. and D.M.L. conceived the experiments and took preliminary data. C.J.B. and V.M.S. designed and performed the experiments described here, analysed data and produced the figures. C.J.B. and D.M.L. wrote the manuscript, which all authors discussed.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.J.B. ([c.ballance@physics.ox.ac.uk](mailto:c.ballance@physics.ox.ac.uk)).

## METHODS

**Ion crystal order.** The  $^{40}\text{Ca}^+ - ^{43}\text{Ca}^+$  ion crystal ordering is kept constant during the experiments to control systematic errors. The principal error that would arise if the ion order were not controlled is due to an (undesired) axial magnetic field gradient that causes the magnetic field between the two ions to differ by  $0.18\mu\text{T}$ . This means that the qubit frequencies for the two possible ion orders differ by approximately 5 kHz, which would lead to errors in single-qubit rotations. We measure the frequency of each qubit using slow (typically  $100\mu\text{s}$ ) carrier  $\pi$  pulses, interleaved with the main experimental pulse sequence, which allows us to detect and to correct for both common-mode qubit frequency changes (due to drift in the global magnetic field  $B$ ) and differential changes (due to incorrect ion crystal ordering). If the ion order is wrong, we randomly reorder the crystal until the order is correct with a short period of Doppler heating to melt the crystal, followed by a short period of Doppler cooling.

**Single-qubit phases and light shifts.** Despite the qubits having very different frequencies, no special phase control is needed to implement the entangling gate. The  $^{43}\text{Ca}^+$  qubit phase is tracked by the microwave local oscillator, and the  $^{40}\text{Ca}^+$  qubit phase is tracked by the difference phase of the co-propagating Raman beams, in turn referenced to a radio-frequency local oscillator. The phases of the Raman beams that implement the entangling gate have no relationship to either of the qubit phases. However, the travelling standing wave resulting from the interference of the Raman gate beams also generates an isotope-dependent differential light shift on each qubit with an amplitude that oscillates at the Raman difference frequency  $\delta$ . Over the course of the gate operation this light shift adds phase shifts to the qubits that depend on the (uncontrolled) optical phase difference of the Raman beams. These uncontrolled phase shifts reduce the fidelity of the gate operation. We greatly reduce this light shift error by shaping the turn-on and turn-off of the Raman laser intensities with a characteristic time of  $1\mu\text{s}$ ; we estimate that without this pulse-shaping the light shift would lead to an average gate error of up to 5% (see ref. 27).

We adjust the polarization of each Raman beam individually to null the differential light shift from each single beam on the  $^{40}\text{Ca}^+$  qubit. (The interference of the two gate beams nevertheless gives rise to a polarization modulation which provides the state-dependent force.) Owing to the difference in atomic structure there is a residual light shift on the  $^{43}\text{Ca}^+$  qubit of approximately 0.2% of the light shift for a purely circularly polarized beam of the same intensity and frequency. This small light shift does not cause any significant issues in the experiments reported here; if necessary it could be suppressed further by increasing the Raman detuning at the expense of requiring more Raman beam power.

**State preparation and measurement errors.** To perform individual single-shot qubit readout, we selectively shelve one qubit state of each ion to the  $3\text{D}_{5/2}$  level, then apply the Doppler cooling lasers sequentially in time first for one isotope, then for the other. If an ion was not shelved it fluoresces, and this is detected with a photomultiplier. We simultaneously shelve the two isotopes using a weak 393 nm beam resonant with the  $^{43}\text{Ca}^+ 4\text{S}_{1/2}^{+4} \leftrightarrow 4\text{P}_{3/2}^{+5}$  transition, with a 1.94 GHz sideband (produced by an electro-optic modulator) which drives the  $^{40}\text{Ca}^+ 4\text{S}_{1/2}^{+1/2} \leftrightarrow 4\text{P}_{3/2}^{+3/2}$  transition. An intense 850 nm beam resonant with the  $^{40}\text{Ca}^+ 3\text{D}_{3/2} \leftrightarrow 4\text{P}_{3/2}$  transition makes the shelving for this isotope state-selective, via electromagnetically induced transparency<sup>32</sup>. The  $^{43}\text{Ca}^+$  shelving is state-selective owing to the 3.2 GHz splitting between the two qubit states<sup>33</sup>. Both these shelving processes have a maximum theoretical efficiency of  $\sim 90\%$  due to leakage to  $3\text{D}_{3/2}$  (which for  $^{43}\text{Ca}^+$  could be eliminated using a further 850 nm beam if required<sup>33</sup>), leading to readout errors of  $\bar{\epsilon} \approx 5\%$  when averaged over both qubit states. From independent experiments (similar to those we describe in ref. 20), we estimate the state-preparation error to be approximately 0.1%, which is negligible compared with the readout error.

We measure the readout errors for each qubit state of each isotope, by preparing and measuring each state typically 10,000 times. Since the qubits are measured individually, it is then straightforward to calculate the linear mapping that corrects for the readout errors, provided that they remain constant. The readout errors relevant to the entangling gate experiment (Fig. 2) were measured to be  $\bar{\epsilon}_{40} = 7.7(2)\%$  for  $^{40}\text{Ca}^+$  and  $\bar{\epsilon}_{43} = 4.4(2)\%$  for  $^{43}\text{Ca}^+$  (averaged over both qubit states). Measurements of the readout errors were interleaved with the gate experimental runs, to check for systematic drifts, and were made using the mixed-isotope crystal, to avoid systematic effects associated with ion position. We estimate the systematic uncertainty in determining the readout errors to be approximately 0.1%, less than the statistical error in these measurements. If we did not correct for readout errors, the apparent infidelity in the Bell state would increase by approximately  $\frac{3}{2}(\bar{\epsilon}_{40} + \bar{\epsilon}_{43}) \approx 18\%$ . For the CHSH test, we do not correct for readout errors, but we nevertheless measure them in order to calculate the maximum attainable CHSH parameter ( $S_{\text{max}}$ ).

**Sample size.** No statistical methods were used to predetermine sample size.

32. McDonnell, M. J. *et al.* High-efficiency detection of a single quantum of angular momentum by suppression of optical pumping. *Phys. Rev. Lett.* **93**, 153601 (2004).
33. Myerson, A. H. *et al.* High-fidelity readout of trapped-ion qubits. *Phys. Rev. Lett.* **100**, 200502 (2008).

# Radiative heat transfer in the extreme near field

Kyeongtae Kim<sup>1†\*</sup>, Bai Song<sup>1\*</sup>, Víctor Fernández-Hurtado<sup>2\*</sup>, Woochul Lee<sup>1</sup>, Wonho Jeong<sup>1</sup>, Longji Cui<sup>1</sup>, Dakotah Thompson<sup>1</sup>, Johannes Feist<sup>2</sup>, M. T. Homer Reid<sup>3</sup>, Francisco J. García-Vidal<sup>2,4</sup>, Juan Carlos Cuevas<sup>2</sup>, Edgar Meyhofer<sup>1</sup> & Pramod Reddy<sup>1,5</sup>

**Radiative transfer of energy at the nanometre length scale is of great importance to a variety of technologies including heat-assisted magnetic recording<sup>1</sup>, near-field thermophotovoltaics<sup>2</sup> and lithography<sup>3</sup>. Although experimental advances have enabled elucidation of near-field radiative heat transfer in gaps as small as 20–30 nanometres (refs 4–6), quantitative analysis in the extreme near field (less than 10 nanometres) has been greatly limited by experimental challenges. Moreover, the results of pioneering measurements<sup>7,8</sup> differed from theoretical predictions by orders of magnitude. Here we use custom-fabricated scanning probes with embedded thermocouples<sup>9,10</sup>, in conjunction with new microdevices capable of periodic temperature modulation, to measure radiative heat transfer down to gaps as small as two nanometres. For our experiments we deposited suitably chosen metal or dielectric layers on the scanning probes and microdevices, enabling direct study of extreme near-field radiation between silica–silica, silicon nitride–silicon nitride and gold–gold surfaces to reveal marked, gap-size-dependent enhancements of radiative heat transfer. Furthermore, our state-of-the-art calculations of radiative heat transfer, performed within the theoretical framework of fluctuational electrodynamics, are in excellent agreement with our experimental results, providing unambiguous evidence that confirms the validity of this theory<sup>11–13</sup> for modelling radiative heat transfer in gaps as small as a few nanometres. This work lays the foundations required for the rational design of novel technologies that leverage nanoscale radiative heat transfer.**

Radiative heat transfer in the far field<sup>14</sup>, that is, at gap sizes larger than Wien's wavelength ( $\sim 10\ \mu\text{m}$  at room temperature), is well established. However, near-field radiative heat transfer (NFRHT), where the gap sizes are smaller than Wien's wavelength, remains relatively unexplored<sup>15</sup>. Over the past decade, a series of technical advances have enabled experiments<sup>4–6</sup> for gap sizes as small as 20 nm to study NFRHT and broadly verify the validity of a theoretical framework called fluctuational electrodynamics<sup>11,16–18</sup> for modelling NFRHT. In contrast, recent experiments<sup>7,8</sup> of extreme (e)NFRHT with single-digit nanometre gap sizes ( $< 10\ \text{nm}$ ) between gold (Au) surfaces have questioned the validity of fluctuational electrodynamics and have raised the question of whether additional mechanisms, even of non-radiative origin such as phonon tunnelling<sup>19</sup>, could dominate the heat transfer in this regime. In addition, some newer computational eNFRHT studies<sup>20</sup> on dielectrics have suggested that the local form of fluctuational electrodynamics, in which one assumes the dielectric properties of the media to be local in space, is inadequate for modelling eNFRHT. Yet other computations<sup>21</sup> on dielectrics have asserted that such non-local effects are irrelevant even for gap sizes as small as 1 nm. This disagreement is of great concern because understanding eNFRHT is critical for the development of a range of novel technologies<sup>1–3</sup>. Here, we present experimental and computational results that both demonstrate marked increases in heat fluxes in the extreme near field and establish

the validity of fluctuational electrodynamics for modelling/predicting eNFRHT for dielectric as well as metal surfaces in gap sizes as small as a few nanometres.

Experimental elucidation of radiative heat transfer across few-nanometre-sized gaps is exceedingly difficult, owing to numerous technical challenges in creating and stably maintaining such gaps while simultaneously measuring minute (pW) heat currents across them. One key innovation used in this work to overcome the technical challenges was to leverage highly sensitive, custom-fabricated probes with embedded Au–Cr thermocouples (Fig. 1a–c), called scanning thermal microscopy (SThM) probes<sup>9</sup>. The SThM probes were fabricated by deposition of multiple metal and dielectric layers to create a nanoscopically small Au–Cr thermocouple at the very end of the tip. Our probes were optimized to have both a high thermal resistance<sup>22</sup> ( $R_p \approx 10^6\ \text{K W}^{-1}$ ) and stiffness<sup>9</sup> ( $> 4\ \text{N m}^{-1}$ ), and were coated with a desired dielectric (silica ( $\text{SiO}_2$ ) or silicon nitride ( $\text{SiN}$ )) or metal (Au) layer. The resulting probes have tip diameters ranging from 350 nm to 900 nm (for details see Fig. 1b and Supplementary Figs 1–3).

The basic strategy for quantifying NFRHT is to record the tip temperature, via the embedded nanoscale thermocouple, which rises in proportion to the radiative heat flow when the tip is displaced towards a heated substrate. To eliminate conductive and convective heat transfer and to remove any water adsorbed to the surfaces, all measurements were performed in an ultra-high vacuum (UHV) using a modified scanning probe microscope (RHK UHV 7500) housed in an ultra-low-noise facility (see Supplementary Information). In performing the measurements, the substrate is heated to an elevated temperature ( $T_S = 425\ \text{K}$ ) while the SThM probe, mounted in the scanner of the scanning probe microscope, is connected to a thermal reservoir maintained at a temperature  $T_R = 310\ \text{K}$ . The spatial separation between the probe and the substrate is reduced at a constant rate of  $0.5\ \text{nm s}^{-1}$  from a gap size of 50 nm until probe–substrate contact. During this process the temperature difference between the tip ( $T_p$ ) and the reservoir ( $T_R$ ),  $\Delta T_p = T_p - T_R$ , is monitored (see Supplementary Information) via the embedded thermocouple, while the deflection of the cantilever is concurrently measured optically via an incident laser (Fig. 1a).

A typical deflection trace for a  $\text{SiO}_2$ -coated tip approaching a  $\text{SiO}_2$ -coated surface is shown in Fig. 2a. From the deflection trace it is apparent that the gap size can be controllably reduced to values as small as  $\sim 2\ \text{nm}$ , below which the tip rapidly 'snaps' towards the substrate and makes contact (see Supplementary Information). This instability is created by attractive forces between the tip and the substrate that arise owing to Casimir and/or electrostatic forces. Figure 2a shows the simultaneously measured  $\Delta T_p$ , which represents the sudden increase in temperature that occurs when the tip snaps into the substrate. This rapid increase in tip temperature ( $\sim 2\ \text{K}$ ) upon mechanical contact is due to heat conduction, via the solid–solid contact, from the hot substrate (425 K) to the tip of the SThM probe, the temperature of

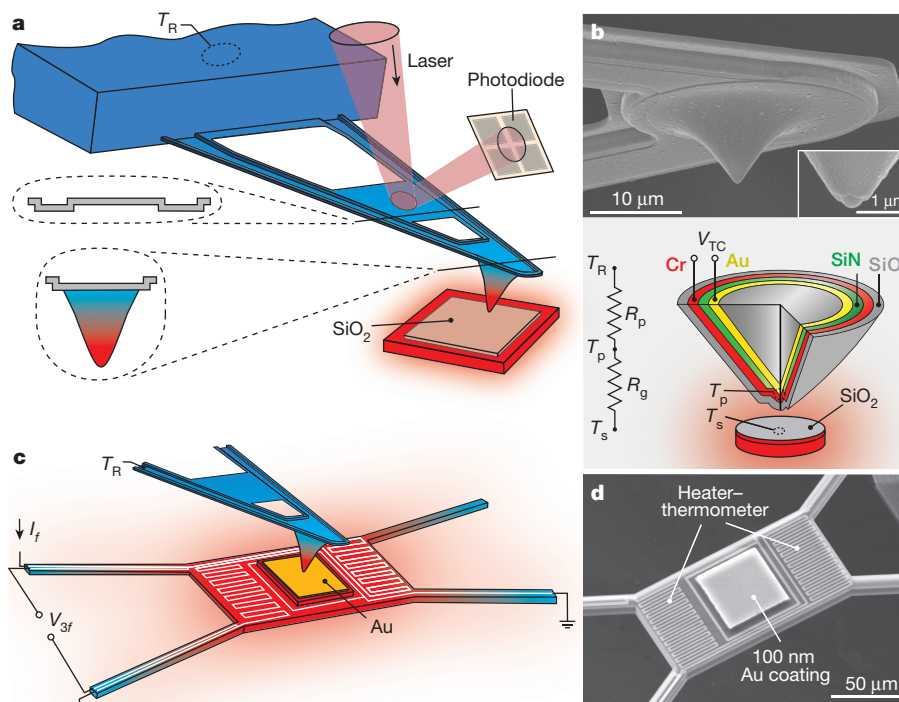
<sup>1</sup>Department of Mechanical Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA. <sup>2</sup>Departamento de Física Teórica de la Materia Condensada and Condensed Matter Physics Center (IFIMAC), Universidad Autónoma de Madrid, Madrid 28049, Spain. <sup>3</sup>Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

<sup>4</sup>Donostia International Physics Center (DIPC), Donostia/San Sebastián 20018, Spain. <sup>5</sup>Department of Materials Science and Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA.

<sup>†</sup>Present address: Department of Mechanical Engineering and Robotics, Incheon National University, Incheon 22012, South Korea.

\*These authors contributed equally to this work.





**Figure 1 | Experimental set-up and SEM images of SThM probes and suspended microdevices.**

**a**, Schematic of the experimental set-up, in which an SThM probe is in close proximity to a heated substrate (insets show cross-sections of the SThM probe). The scenario for  $\text{SiO}_2$  measurements is shown (the coating on the substrate is replaced with SiN and Au in other experiments). **b**, SEM image (top) of a SThM probe. The inset shows an SEM image of the hemispherical probe tip, which features an embedded Au–Cr thermocouple from which the thermoelectric voltage  $V_{TC}$  is measured. The bottom panel illustrates a schematic cross-section for a  $\text{SiO}_2$ -coated probe used in  $\text{SiO}_2$  measurements. For SiN and

Au measurements, the outer  $\text{SiO}_2$  coating is appropriately substituted as explained in Supplementary Information. A resistance network that describes the thermal resistance of the probe ( $R_p$ ) and the vacuum gap ( $R_g = (G_{\text{eNFRHT}})^{-1}$ ), as well as the temperatures of the substrate ( $T_s$ ), tip ( $T_p$ ) and reservoir ( $T_R$ ) is also shown. **c**, Schematic showing the measurement scheme used for high-resolution eNFRHT measurements of Au–Au. The amplitude of the supplied sinusoidal electric current is  $I_f$ , the sinusoidal temperature oscillations at  $2f$  are related to the voltage output  $V_{3f}$ . **d**, SEM image of the suspended microdevice featuring the central region coated with Au and a serpentine Pt heater–thermometer.

which is  $\sim 400$  K (heating by the incident laser results in an elevated temperature).

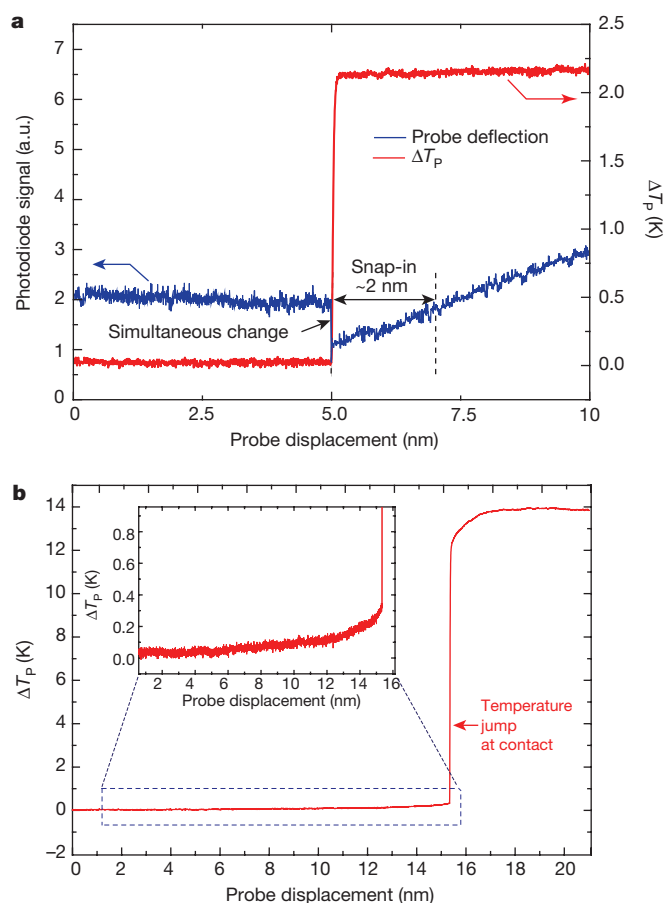
The tight temporal correlation between the mechanical snap-in and the temperature jump of the probe makes it possible to identify tip–substrate contact solely on the basis of temperature signals. In Fig. 2b, the recorded tip temperature is shown as a probe approaches a heated substrate with the laser beam turned off. The recorded temperature signals with and without laser tracking are basically identical (Fig. 2a, b), except that the magnitude of the jump reflects the tip–substrate temperature difference with and without laser excitation. Thus, mechanical contact can be readily detected from the robust temperature jump without laser excitation, thereby avoiding probe heating and laser interference effects. Therefore, we performed all experiments by first estimating the snap-in distance using the optical scheme and subsequently turning the laser off to perform eNFRHT measurements (see Supplementary Information for the measurement of gap size and snap-in distance).

To determine the gap ( $d$ )-dependent near-field radiative conductance ( $G_{\text{eNFRHT}}$ ), we measured  $\Delta T_p$  and directly estimated  $G_{\text{eNFRHT}}$  from  $G_{\text{eNFRHT}}(d) = \Delta T_p / [R_p(T_s - T_R - \Delta T_p)]$ , where  $R_p$  is the thermal resistance of the probe, which was experimentally determined as described in Supplementary Information (Supplementary Fig. 7) to be  $1.6 \times 10^6 \text{ K W}^{-1}$  and  $1.3 \times 10^6 \text{ K W}^{-1}$  for the  $\text{SiO}_2$ - and SiN-coated probes, respectively. The measured conductance of the gaps for  $\text{SiO}_2$  and SiN surfaces is shown in Fig. 3a and b, respectively. It can be seen that  $G_{\text{eNFRHT}}$  increases monotonically until the probe snaps into contact (gap size at snap-in is  $\sim 2$  nm for both  $\text{SiO}_2$  and SiN measurements; see Supplementary Information and Supplementary Fig. 6). Furthermore, it can be seen that the eNFRHT is larger for experiments performed with  $\text{SiO}_2$ . These measurements represent

the first observation of eNFRHT in single-digit nanometre-sized gaps between dielectric surfaces. We compared these results to our computational predictions based on fluctuational electrodynamics, assuming local-dielectric properties (see details later), and found very good agreement (blue lines in Fig. 3a, b).

The remarkable agreement between eNFRHT measurements and computational predictions raises important questions with regards to recent experiments<sup>7</sup> investigating eNFRHT between Au surfaces, which suggested strong disagreements ( $\sim 500$ -fold) between predictions of fluctuational electrodynamics and the results of experiments. One may wonder if the good agreement reported above is unique to eNFRHT between polar dielectric materials. To answer this question unambiguously, we performed additional eNFRHT measurements with Au-coated probes and substrates. The measured conductance in these experiments is shown in Fig. 3c. It can be seen that the measured  $G_{\text{eNFRHT}}$  with decreasing gap size remains comparable to the noise floor of  $\sim 220 \text{ pW K}^{-1}$  for Au-coated probes at an applied temperature differential of  $\sim 115$  K (see Supplementary Information) and is much smaller than that observed for polar dielectrics. These measurements set an upper bound of  $\sim 250 \text{ pW K}^{-1}$  for  $G_{\text{eNFRHT}}$  in our Au–Au experiments. This result is particularly surprising because previous studies that used probes with smaller diameters and lower thermal resistances<sup>7,23</sup> ( $(23\text{--}54) \times 10^3 \text{ K W}^{-1}$  and  $\sim 10^6 \text{ K W}^{-1}$ , implying a lower sensitivity than our probes) reported conductances  $> 40 \text{ nW K}^{-1}$ , which are at least two orders of magnitude larger than conductances measured by us and predicted by theory.

To resolve this contradiction we needed to improve the resolution of our conductance measurements by more than an order of magnitude (see Supplementary Information and Supplementary



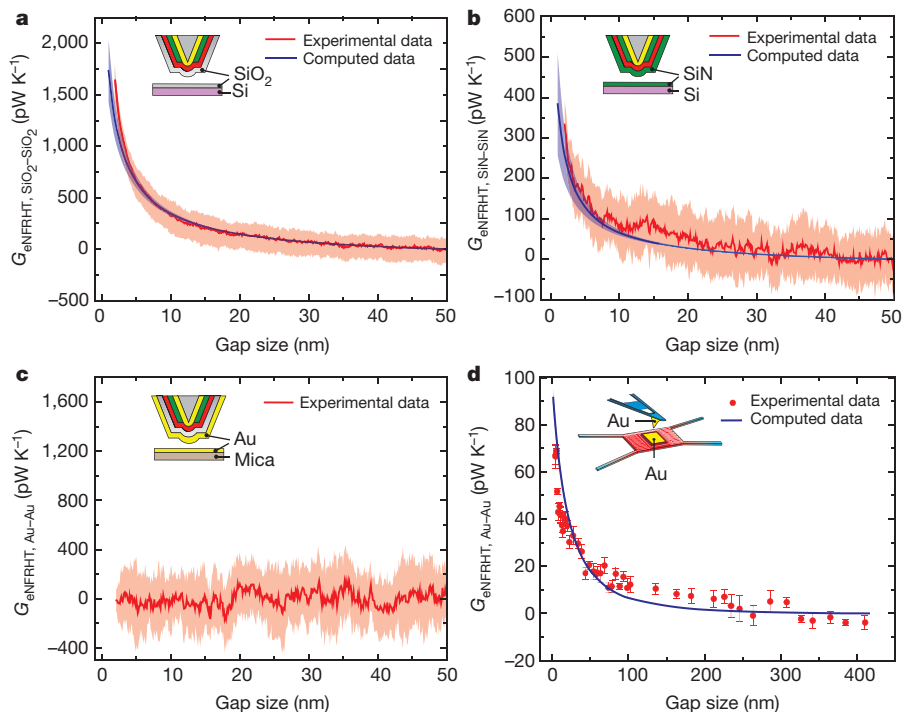
**Figure 2 | Detection of mechanical contact from deflection and temperature signals.** **a**, Data from an experiment in which a SiO<sub>2</sub>-coated probe at about 400 K (heated by the incident laser) is displaced towards a heated SiO<sub>2</sub> substrate at 425 K. The deflection of the scanning probe (blue), reported in arbitrary units (a.u.), and rise in temperature of probe,  $\Delta T_P$  (red), are shown. The sudden decrease in the deflection signal due to snap-in coincides with a simultaneous increase in the tip temperature due to conduction of heat from the hot substrate to the cold tip, clearly showing that contact can be readily detected by the large temperature jump. The snap-in distance is seen to be  $\sim 2$  nm. **b**, Measured  $\Delta T_P$  when an unheated probe (310 K, laser turned off) is displaced towards the substrate. A sudden increase in the tip temperature is seen when the cold tip contacts the substrate. Inset shows the increase in the tip temperature due to eNFRHT.

Fig. 8 for details). This was accomplished by using a new microdevice (see Fig. 1c, d and Supplementary Figs 4, 5, 9, 10 for details of device fabrication and characterization) that features a suspended island whose temperature can be readily modulated at  $f = 18$  Hz (see Supplementary Information). Sinusoidal electric currents (9 Hz) supplied to the embedded electrical heater resulted in sinusoidal temperature oscillations at the second harmonic with amplitude ( $\Delta T_{S,f=18\text{ Hz}}$ ) that was accurately measured using a lock-in technique<sup>6,24</sup> (see Supplementary Information). To characterize eNFRHT, we positioned a Au-coated SThM probe (30 nm Au thickness) in close proximity to the surface of the microfabricated device, which features a suspended region that is  $50\text{ }\mu\text{m} \times 50\text{ }\mu\text{m}$  large and was coated with 100 nm of Au. The amplitude of temperature modulation of the probe ( $\Delta T_{P,f=18\text{ Hz}}$ ), due to eNFRHT, was measured at various gap sizes (see Supplementary Information) in a bandwidth of 0.78 mHz. Given the low noise in this bandwidth it was possible to resolve temperature changes as small as  $\sim 20\text{ }\mu\text{K}$ , which corresponds to a conductance noise floor of  $\sim 6\text{ pW K}^{-1}$ , when  $\Delta T_{S,f=18\text{ Hz}}$  is 5 K (see Supplementary Information section 7 for details of the noise characterization). The measured  $\Delta T_{P,f=18\text{ Hz}}$  values were

used to estimate  $G_{\text{eNFRHT}}$  (Fig. 3d) via:  $G_{\text{eNFRHT}}(d) = \Delta T_{P,f=18\text{ Hz}} / [R_{\text{P,Au}}(\Delta T_{S,f=18\text{ Hz}} - \Delta T_{P,f=18\text{ Hz}})]$ , where  $R_{\text{P,Au}} = 0.7 \times 10^6\text{ K W}^{-1}$  is the thermal resistance of the Au-coated probe (see Supplementary Information and Supplementary Fig. 7). The smallest gap size at which measurements could be accomplished is  $\sim 3$  nm and is limited by both snap-in and deflections of the microdevice due to periodic thermal expansion resulting from bimaterial effects (see Supplementary Fig. 11). The measured  $G_{\text{eNFRHT}}$  (Fig. 3d) is indeed much smaller than that obtained with SiO<sub>2</sub> (Fig. 3a) and SiN (Fig. 3b) films. In contrast to previous experiments<sup>7</sup>, our measured  $G_{\text{eNFRHT}}$  for Au–Au surfaces is in excellent agreement with the predictions of fluctuational electrodynamics (solid line in Fig. 3d).

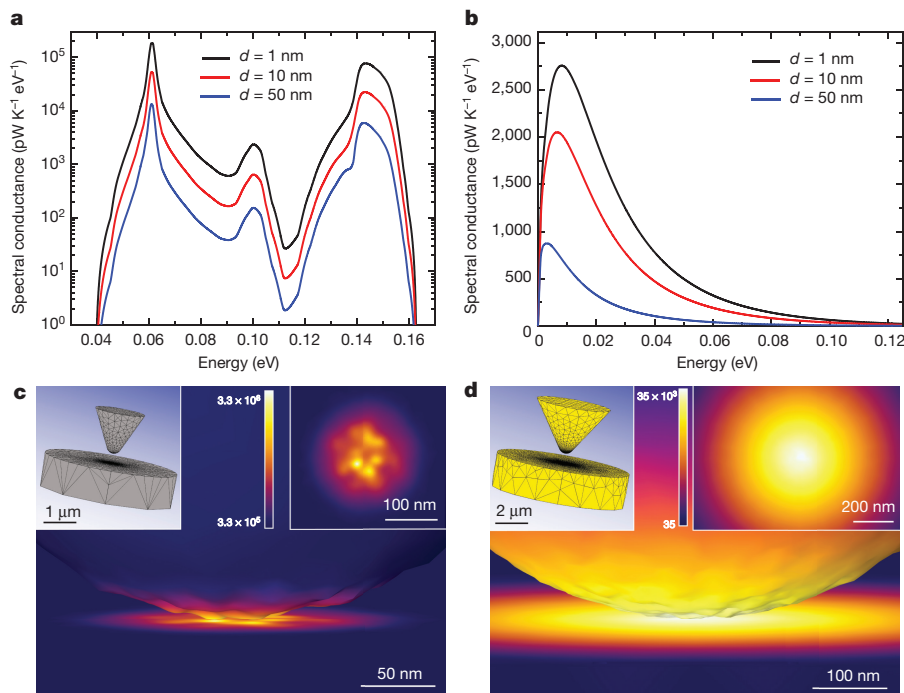
To obtain insight into our experimental results, we used a fluctuating-surface-current formulation of the radiative heat transfer problem<sup>13,25</sup> combined with the boundary element method, as implemented by us in the SCUFF-EM solver<sup>26</sup>. This allows NFRHT calculations between bodies of arbitrary shape and provides numerically exact results within the framework of fluctuational electrodynamics in the local approximation<sup>13,25</sup>. For our calculations, we characterized the dielectric function for SiN, whereas the dielectric functions for SiO<sub>2</sub> and Au were taken from previous work (see Supplementary Information section 12 and Supplementary Fig. 12). To simulate our experiments accurately, we considered the tip–substrate geometries shown in the left insets of Fig. 4c, d. Here, the tip has a conical shape and ends in a spherical cap whose radius was obtained from scanning electron microscope (SEM) images of the probes (see Supplementary Figs 1–3). In our simulations, we included sufficiently large areas of the probe's conical part and the substrate such that the results do not depend on their finite size (see Supplementary Information section 14 and Supplementary Fig. 13). To maintain high fidelity to the experimental conditions, we also accounted for the small roughness of our probes by including random Gaussian-correlated noise in the tip profile (Fig. 4c, d). More precisely, the maximum protrusion height on the tip and the correlation length between protrusions were chosen to be 10 nm and 17 nm, respectively, on the basis of the surface characteristics observed in the SEM images (Supplementary Figs 1–3). We investigated the effect of surface roughness by computing  $G_{\text{eNFRHT}}$  for every material from 15 different tip–substrate ensembles with roughness profiles generated as described earlier. The computational results for the different materials are presented in Fig. 3a, b, d. As pointed out earlier, we indeed find very good agreement between computation and experiment without any adjustable parameters.

To elucidate the underlying physical mechanism and explain the differences in eNFRHT between different material combinations, we computed the spectral conductance (heat conductance per unit of energy) for several gap sizes as shown in Fig. 4a, b for SiO<sub>2</sub> and Au, respectively (see Supplementary Fig. 14 for SiN results). In Fig. 4a, one can see that the dominant contributions to the spectral conductance of SiO<sub>2</sub> come from two narrow energy ranges centred around  $\sim 0.06\text{ eV}$  and  $\sim 0.14\text{ eV}$ , which correspond to the energies of the transverse optical phonons of SiO<sub>2</sub>. This strongly suggests that for SiO<sub>2</sub>, eNFRHT is dominated by surface phonon polaritons (SPhPs), as previously found for larger gaps<sup>6,27,28</sup>. In turn, this explains the marked decrease in heat transfer as the gap size increases, which is a consequence of the rapid decrease in the number of available surface electromagnetic modes for radiation to tunnel across the vacuum gap. In contrast, eNFRHT for Au exhibits a rather broad spectral conductance that decays more slowly with gap size (Fig. 4b). This slow decay is reminiscent of the situation encountered in a plate–plate geometry<sup>29</sup> where NFRHT is dominated by frustrated internal reflection modes, that is, by modes that are evanescent in the vacuum gap but are propagating inside the Au tip and substrate whose contribution saturates for gaps below the skin depth<sup>29</sup>, which for Au is around 25 nm. This naturally explains the weaker dependence of eNFRHT on gap size observed in our Au–Au measurements. The fundamental difference in eNFRHT between dielectrics and metals is also apparent from the computed Poynting-flux



**Figure 3 | Measured extreme near-field thermal conductances for dielectric and metal surfaces.** **a**, Measured near-field radiative conductance between a SiO<sub>2</sub>-coated probe (310 K) and a SiO<sub>2</sub> substrate at 425 K. The red solid line shows the average conductance from 15 independent measurements, the light red band represents the standard deviation. The blue solid line shows the average of the computed radiative conductance for 15 different tips with stochastically chosen roughness profiles (root-mean-squared roughness of  $\sim 10$  nm) and a tip diameter (450 nm) obtained from SEM images of the probe.

The blue shaded region represents the standard deviation in the calculated data. **b**, **c**, Same as **a**, but for SiN-SiN and Au-Au, respectively. The tip diameter is 350 nm for the SiN-coated tip. Computed results are not included for Au-Au. **d**, Near-field conductance from experiments with a Au-coated probe and a suspended microdevice. Red dots represent the average from 10 different measurements (temperature periodically modulated at 18 Hz); the error bars represent the standard deviation. The blue solid line represents the computed conductance (tip diameter is 900 nm).



**Figure 4 | Spectral conductance and spatial distribution of the Poynting flux.** **a**, Spectral conductance as a function of energy for a SiO<sub>2</sub> tip-substrate geometry for three different gap sizes. The tip diameter is 450 nm, and the reservoir temperatures are 310 K for the tip and 425 K for the substrate. Notice the logarithmic scale in the vertical axis. **b**, Same as **a**, but for Au. In this case, the tip radius is 450 nm, and the tip and substrate temperatures are 300 K and 301 K, respectively. **c**, Surface-contour plot showing the spatial distribution of the Poynting-flux pattern on the

surface of the bodies for the SiO<sub>2</sub> tip-substrate geometry corresponding to that in **a** with a gap of 1 nm. The colour scale is in units of  $W (K eV m^2)^{-1}$  and the plot was computed at an energy of 61 meV, which corresponds to the maximum of the spectral conductance. The right inset shows the corresponding surface heat flux on the substrate; the left inset displays the whole tip-substrate geometry simulated, including the mesh used in the calculations. **d**, Same as **c**, but for Au. In this case the surface-contour plot was computed at 9 meV, the maximum of the spectral conductance.



patterns on the surfaces (Fig. 4c, d), which show that eNFRHT in the SiO<sub>2</sub> case is much more concentrated in the tip apex than it is in the Au case. This difference reflects the fact that in a polar dielectric, such as SiO<sub>2</sub>, eNFRHT has a very strong distance dependence due to the excitation of SPhPs with a penetration depth comparable to the gap size<sup>6</sup>. Given these differences between metals and dielectrics, it is not surprising that Au–Au eNFRHT is relatively insensitive to small surface roughness (see Supplementary Fig. 15). For this reason, the large differences between our results for Au and those of previous work<sup>7,8</sup>, which disagree with the predictions of fluctuational electrodynamics, cannot be attributed to differences in the surface roughness. Our computational results, when compared with our experimental data, provide unambiguous evidence that fluctuational electrodynamics accurately describes eNFRHT.

We note that the results presented here provide the first experimental evidence—to our knowledge—for extremely large enhancements of radiative heat transfer in the extreme near field between both dielectric and metal surfaces. Furthermore, our results establish the fundamental validity of fluctuational electrodynamics in modelling eNFRHT and NFRHT. The technical advances described in this work are key to systematically investigating eNFRHT phenomena in a variety of materials and nanostructures, and provide critical information that complements insights that can be obtained by other near-field techniques<sup>30,31</sup>. Knowledge gained from such studies will be critical to the development of future technologies that leverage nanoscale radiative heat transfer<sup>32</sup>.

Received 10 August; accepted 1 October 2015.

Published online 7 December 2015.

- Challener, W. A. *et al.* Heat-assisted magnetic recording by a near-field transducer with efficient optical energy transfer. *Nature Photon.* **3**, 220–224 (2009).
- Basu, S., Zhang, Z. M. & Fu, C. J. Review of near-field thermal radiation and its application to energy conversion. *Int. J. Energy Res.* **33**, 1203–1232 (2009).
- Pendry, J. B. Radiative exchange of heat between nanostructures. *J. Phys. Condens. Matter* **11**, 6621–6633 (1999).
- Shen, S., Narayanaswamy, A. & Chen, G. Surface phonon polaritons mediated energy transfer between nanoscale gaps. *Nano Lett.* **9**, 2909–2913 (2009).
- Rousseau, E. *et al.* Radiative heat transfer at the nanoscale. *Nature Photon.* **3**, 514–517 (2009).
- Song, B. *et al.* Enhancement of near-field radiative heat transfer using polar dielectric thin films. *Nature Nanotechnol.* **10**, 253–258 (2015).
- Kittel, A. *et al.* Near-field heat transfer in a scanning thermal microscope. *Phys. Rev. Lett.* **95**, 224301–224304 (2005).
- Worbes, L., Hellmann, D. & Kittel, A. Enhanced near-field heat flow of a monolayer dielectric island. *Phys. Rev. Lett.* **110**, 134302 (2013).
- Kim, K., Jeong, W., Lee, W. & Reddy, P. Ultra-high vacuum scanning thermal microscopy for nanometer resolution quantitative thermometry. *ACS Nano* **6**, 4248–4257 (2012).
- Lee, W. *et al.* Heat dissipation in atomic-scale junctions. *Nature* **498**, 209–212 (2013).
- Rytov, S. M. *Theory of Electric Fluctuations and Thermal Radiation* (Air Force Cambridge Research Center, 1953).
- Joulain, K., Mulet, J.-P., Marquier, F., Carminati, R. & Greffet, J.-J. Surface electromagnetic waves thermally excited: radiative heat transfer, coherence properties and Casimir forces revisited in the near field. *Surf. Sci. Rep.* **57**, 59–112 (2005).
- Rodriguez, A. W., Reid, M. T. H. & Johnson, S. G. Fluctuating-surface-current formulation of radiative heat transfer: theory and applications. *Phys. Rev. B* **88**, 054305 (2013).
- Planck, M. & Masius, M. *The Theory of Heat Radiation* (P. Blakiston Son & Co, 1914).
- Song, B., Fiorino, A., Meyhofer, E. & Reddy, P. Near-field radiative thermal transport: From theory to experiment. *AIP Adv.* **5**, 053503 (2015).
- Rytov, S. M., Kravtsov, Y. A. & Tatarskii, V. I. *Principles of Statistical Radiophysics* (Springer, 1989).
- Polder, D. & Hove, M. A. V. Theory of radiative heat transfer between closely spaced bodies. *Phys. Rev. B* **4**, 3303–3314 (1971).
- Shen, S., Mavrokefalos, A., Sambegoro, P. & Chen, G. Nanoscale thermal radiation between two gold surfaces. *Appl. Phys. Lett.* **100**, 233114 (2012).
- Alteder, I., Voevodin, A. A. & Roy, A. K. Vacuum phonon tunneling. *Phys. Rev. Lett.* **105**, 166101 (2010).
- Singer, F., Ezzahri, Y. & Joulain, K. Near field radiative heat transfer between two nonlocal dielectrics. *J. Quant. Spectrosc. Radiat. Transf.* **154**, 55–62 (2015).
- Chiloyan, V., Garg, J., Estarjani, K. & Chen, G. Transition from near-field thermal radiation to phonon heat conduction at sub-nanometre gaps. *Nature Commun.* **6**, 6755 (2015).
- Kim, K. *et al.* Quantification of thermal and contact resistances of scanning thermal probes. *Appl. Phys. Lett.* **105**, 203107 (2014).
- Wischnath, U. F., Welker, J., Munzel, M. & Kittel, A. The near-field scanning thermal microscope. *Rev. Sci. Instrum.* **79**, 073708 (2008).
- Sadat, S., Meyhofer, E. & Reddy, P. Resistance thermometry-based picowatt-resolution heat-flow calorimeter. *Appl. Phys. Lett.* **102**, 163110–163113 (2013).
- Rodriguez, A. W., Reid, M. T. H. & Johnson, S. G. Fluctuating-surface-current formulation of radiative heat transfer for arbitrary geometries. *Phys. Rev. B* **86**, 220302 (2012).
- Reid, M. T. H. & Johnson, S. G. Efficient computation of power, force and torque in BEM scattering calculations. *IEEE Trans. Antenn. Propag.* **63**, 3588–3598 (2015).
- Mulet, J. P., Joulain, K., Carminati, R. & Greffet, J. J. Enhanced radiative heat transfer at nanometric distances. *Microscale Therm. Eng.* **6**, 209–222 (2002).
- Mulet, J. P., Joulain, K., Carminati, R. & Greffet, J. J. Nanoscale radiative heat transfer between a small particle and a plane surface. *Appl. Phys. Lett.* **78**, 2931–2933 (2001).
- Chapuis, P. O., Volz, S., Henkel, C., Joulain, K. & Greffet, J. J. Effects of spatial dispersion in near-field radiative heat transfer between two parallel metallic surfaces. *Phys. Rev. B* **77**, 035431 (2008).
- Jones, A. C. & Raschke, M. B. Thermal infrared near-field spectroscopy. *Nano Lett.* **12**, 1475–1481 (2012).
- De Wilde, Y. *et al.* Thermal radiation scanning tunnelling microscopy. *Nature* **444**, 740–743 (2006).
- Otey, C. R., Lau, W. T. & Fan, S. H. Thermal rectification through vacuum. *Phys. Rev. Lett.* **104**, 154301 (2010).

Supplementary Information is available in the online version of the paper.

**Acknowledgements** P.R. acknowledges support from US Department of Energy Basic Energy Sciences through a grant from the Scanning Probe Microscopy Division under award no. DE-SC0004871 (fabrication of scanning thermal probes). E.M. and P.R. acknowledge support from the Army Research Office under grant W911NF-12-1-0612 (fabrication of microdevices). P.R. acknowledges support from the Office of Naval Research under grant award no. N00014-13-1-0320 (instrumentation). E.M. and P.R. acknowledge support from the National Science Foundation under grant CBET 1235691 (thermal characterization). J.C.C. acknowledges financial support from the Spanish Ministry of Economy and Competitiveness (MINECO) (contract no. FIS2014-53488-P) and the Comunidad de Madrid (contract no. S2013/MIT-2740) and V.F.-H. from “la Caixa” Foundation. F.J.G.-V. and J.F. acknowledge support from the European Research Council (ERC-2011-AdG Proposal No. 290981), the European Union Seventh Framework Programme (FP7-PEOPLE-2013-CIG-618229), and the Spanish MINECO (MAT2011-28581-C02-01 and MAT2014-53432-C5-5-R). The authors acknowledge the Lurie Nanofabrication Facility for facilitating the nanofabrication of devices.

**Author Contributions** The work was conceived by P.R., E.M., F.J.G.-V. and J.C.C. The experiments were performed by K.K., W.L., L.C. and B.S. under the supervision of E.M. and P.R. The devices were designed, fabricated and characterized by K.K., W.J., D.T. and B.S. Characterization of dielectric properties was performed by B.S. Modelling was performed by V.F.-H., J.F. and B.S. (with inputs from M.T.H.R.) under the supervision of F.J.G.-V. and J.C.C. The manuscript was written by J.C.C., E.M. and P.R. with comments and inputs from all authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.C.C. ([juancarlos.cuevas@uam.es](mailto:juancarlos.cuevas@uam.es)) or E.M. ([meyhofer@umich.edu](mailto:meyhofer@umich.edu)) or P.R. ([pramodr@umich.edu](mailto:pramodr@umich.edu)).

# Self-shaping of oil droplets via the formation of intermediate rotator phases upon cooling

Nikolai Denkov<sup>1</sup>, Slavka Tcholakova<sup>1</sup>, Ivan Lesov<sup>1</sup>, Diana Cholakova<sup>1</sup> & Stoyan K. Smoukov<sup>2</sup>

Revealing the chemical and physical mechanisms underlying symmetry breaking and shape transformations is key to understanding morphogenesis<sup>1</sup>. If we are to synthesize artificial structures with similar control and complexity to biological systems, we need energy- and material-efficient bottom-up processes to create building blocks of various shapes that can further assemble into hierarchical structures. Lithographic top-down processing<sup>2</sup> allows a high level of structural control in microparticle production but at the expense of limited productivity. Conversely, bottom-up particle syntheses<sup>3–8</sup> have higher material and energy efficiency, but are more limited in the shapes achievable. Linear hydrocarbons are known to pass through a series of metastable plastic rotator phases before freezing<sup>9,10</sup>. Here we show that by using appropriate cooling protocols, we can harness these phase transitions to control the deformation of liquid hydrocarbon droplets and then freeze them into solid particles, permanently preserving their shape. Upon cooling, the droplets spontaneously break their shape symmetry several times, morphing through a series of complex regular shapes owing to the internal phase-transition processes. In this way we produce particles including micrometre-sized octahedra, various polygonal platelets, O-shapes, and fibres of submicrometre diameter, which can be selectively frozen into the corresponding solid particles. This mechanism offers insights into achieving complex morphogenesis from a system with a minimal number of molecular components.

We illustrate the capabilities of this new approach by using droplets of different linear hydrocarbons with 14–20 carbon atoms (namely from tetradecane to eicosane). The alkanes were pre-dispersed as droplets in 1.5 wt% aqueous surfactant solution, which can afterwards be transformed into a variety of solid particles with different shapes through the choice of appropriate surfactants and controlled cooling rates (Fig. 1 and Extended Data Fig. 1). Figure 2 shows how the choice of surfactant can influence the particle shape and aspect ratio. Upon freezing, many of the thin, high-aspect-ratio structures obtained develop a puncture hole in their interior, owing to the volumetric shrinkage accompanying solidification.

Our experiments show that the drop-shape transformations and the final shape of the frozen particles depend most on three factors: surfactant type, cooling rate, and initial droplet size. We outline the main effects of these control factors and, afterwards, we explain the basic mechanism of the drop shape evolution.

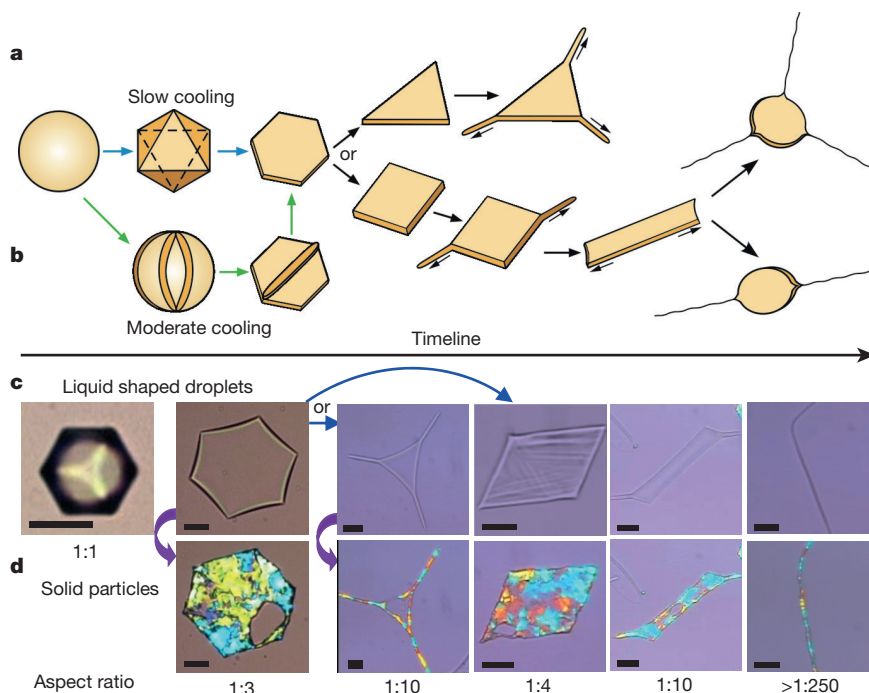
Surfactants are amphiphilic molecules with a hydrophilic head-group (ionic or non-ionic) and hydrophobic alkyl chain. While only two of our surfactants were ultra-pure ( $C_{16}H_{33}N(CH_3)_3Br$  (CTAB) and  $C_{14}H_{29}SO_4Na$  with purity >99%), our experiments with more than ten surfactants of all types (anionic, cationic and non-ionic) showed the same general sequence of shape transformations (Fig. 1). These transformations occurred only when the surfactant chain length was similar to or longer than the length of the hydrocarbon molecules in the droplets. Such long-chain surfactants can freeze in the adsorption

layer on the drop–water interface, before the freezing of the alkane in the droplet interior<sup>10</sup>, and thus have a critical role in the formation of drops with non-spherical shapes. The use of surfactants with shorter chain lengths led to drops freezing into spherical solid particles, without any peculiar shape transformations.

The rate of cooling is another crucial factor in the observed phenomenon. Upon slow and moderate cooling rates (below about  $4\text{ K min}^{-1}$ ), the spherical hydrocarbon drops undergo a series of shape transformations. Figure 1 illustrates the case of hexadecane drops in water containing 1.5 wt% of the non-ionic surfactant  $C_{16}H_{33}(CH_2CH_2O)_{20}OH$  (Brij 58). Initially, the spherical drops transform into regular octahedra (regular polyhedra whose surface is shaped by eight triangular facets) that then transform into flat platelets with a hexagonal base. Upon further cooling, these hexagons transform either into triangular or tetragonal platelets, the ratio of which depends on the surfactant, the cooling rate, and the initial size of the droplets. Subsequently, rod-like asperities with diameters of around  $5\text{ }\mu\text{m}$  appear and grow into long filaments from the platelet tips. Finally, if the cooling is sufficiently slow (less than  $0.5\text{ K min}^{-1}$ ), these asperities elongate further to form very thin fibres with diameters of around  $0.5\text{ }\mu\text{m}$ . When the cooling rates were varied between  $0.01$  and  $2\text{ K min}^{-1}$ , each transformation took between 30 s and several minutes (see Supplementary Video 1). Depending on the cooling rate, the drops freezing into solid particles occurred at different stages of this evolution path—slower cooling led to freezing at a later stage. Thus, using an appropriate intermediate cooling rate, we could transform an intermediate drop shape into a solid frozen particle with the same shape (Figs 1 and 2; alternatively, one can apply step-acceleration cooling to freeze the deformed drops). For example, using Brij 58 as surfactant, at  $0.2\text{ K min}^{-1}$  cooling with  $30\text{ }\mu\text{m}$  droplets, we obtained  $25 \pm 5\%$  triangles and  $75 \pm 5\%$  rhomboids that evolve into rod-shape particles and then finally into fibres, as determined from observations of over 100 droplets in more than 10 independent experiments. For comparison, when using  $10\text{ }\mu\text{m}$  droplets with Tween 60, at  $0.2\text{ K min}^{-1}$  cooling, we could yield more than 90% rod-shape particles.

Drop size was another important factor. The images shown in Figs 1 and 2 are obtained with drops of initial diameter around  $20\text{ }\mu\text{m}$ . Very similar results were obtained with drops of diameter between 1 and  $50\text{ }\mu\text{m}$ . At low cooling rates ( $0.01$ – $2\text{ K min}^{-1}$ ), both small and large drops evolved in shape when appropriate surfactants were used. However, at higher cooling rates the big drops tended to freeze into spherical solid particles without shape transformations, while the smaller drops readily evolved in shape. Thus we could induce shape transformations of drops with diameter  $1$ – $50\text{ }\mu\text{m}$  in a wide range of cooling rates ( $0.01$ – $2\text{ K min}^{-1}$ ). Note that the small droplets with micrometre size are involved in intensive Brownian motion that, however, does not suppress the shape transformations. The investigation of submicrometre droplets was postponed for a separate study because the observation of their evolution requires much more sophisticated experimental methods.

<sup>1</sup>Department of Chemical and Pharmaceutical Engineering, Faculty of Chemistry and Pharmacy, Sofia University, 1164 Sofia, Bulgaria. <sup>2</sup>Active and Intelligent Materials Laboratory, Department of Materials Science & Metallurgy, University of Cambridge, Cambridge CB3 0FS, UK.



**Figure 1 | Schematic of the shape transformations observed during cooling of emulsion droplets of pure hydrocarbons in water, in the presence of 1.5 wt% surfactant. a,** Under slow cooling the spheres transform consecutively into regular octahedra, hexagonal platelets, triangular (or tetragonal) platelets, platelets with long asperities and,

eventually, thin fibres. **b,** Under moderate cooling the droplet surface corrugates and forms a hexagonal platelet with protrusions, before forming the regular hexagonal platelet. **c,** Hexadecane liquid droplets in Brij 58 solution at different stages. **d,** Solid particles with different shapes, obtained at appropriate cooling rates. Scale bars, 20  $\mu\text{m}$ .

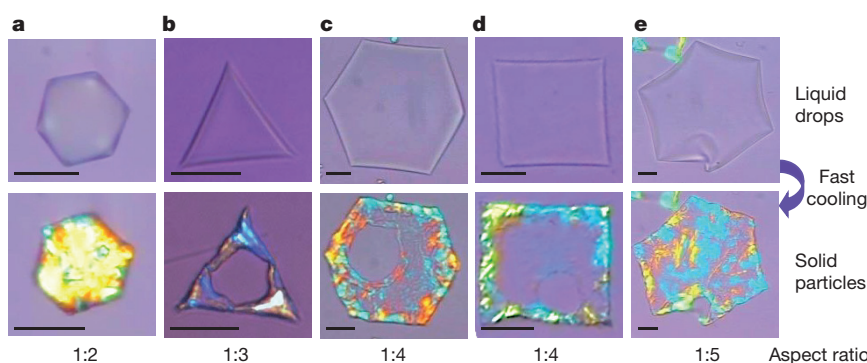
The observed numerous shape transformations are surprisingly governed by a single mechanism. It was deduced from the experimental observations described earlier as follows.

The requirement for a long surfactant chain-length indicates that the shape transformations are triggered by the freezing surfactant adsorption layers, formed on the hydrocarbon–water interface. Numerical estimates described later demonstrate that these adsorption layers do not possess a sufficiently high bending moment to deform the hydrocarbon drops. Therefore, before freezing, the formation of mesomorphic hydrocarbon phases, just inside the surface of the liquid drops, is required to trigger the observed transformations. Indeed, the drop-shape transformations start around the freezing temperature of the bulk hydrocarbon phase transition (18 °C for hexadecane). In this temperature range, linear hydrocarbons form so-called ‘rotator’ mesomorphic phases, which represent a class of plastic phases in which the molecules possess long-range translational order, yet rotate freely around their long axis<sup>9,10</sup>. Owing to their positional long-range order, the rotator phases generate non-isotropic elastic stresses, which are

sufficiently high to deform liquid drops, overcoming their interfacial tension<sup>11</sup>.

Images of deformed drops clearly illustrate the elastic nature of the fluid material confined inside the drops (Fig. 3b and Supplementary Video 2). For non-elastic materials, the long filaments growing from the tips of the triangular platelet shown in Fig. 3a would be unstable and should break into small spherical droplets through so-called ‘Plateau–Rayleigh’ capillary instability, under the action of capillary pressure, which destabilizes cylindrical liquid jets<sup>12</sup>. The angular doughnut-shaped droplets, shown in Fig. 3b, would also be unstable in shape, unless the drop material possesses elasticity to counteract the capillary pressure, forcing the common liquid drops to acquire a spherical shape.

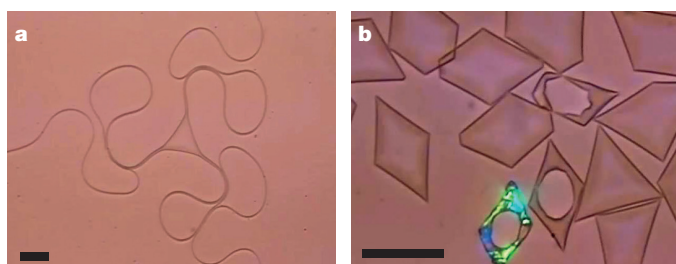
The simplest possible explanation of the observed non-spherical drops could be that the rotator phases, structured inside the entire volume of the drop interior, create high elastic stresses that deform the drop surface. However, some of our results contradict this simplest explanation. For example, the optical microscopy observations of the



**Figure 2 | Choice of surfactant and cooling rate can determine particle shape and aspect ratio. a–e,** Microscope images of deformed liquid droplets (top) and of the solid particles obtained from these droplets (bottom), upon cooling of hexadecane-in-water emulsions, stabilized by

various surfactants: **a, b,** Tween 60; **c, d,** Tween 40; **e,** Brij 78. The numbers indicate the aspect ratio. The aspect ratio of the various shapes along the drop evolution (compare with Fig. 1) depends on the specific surfactant used to stabilize the drops. Scale bars, 20  $\mu\text{m}$ .

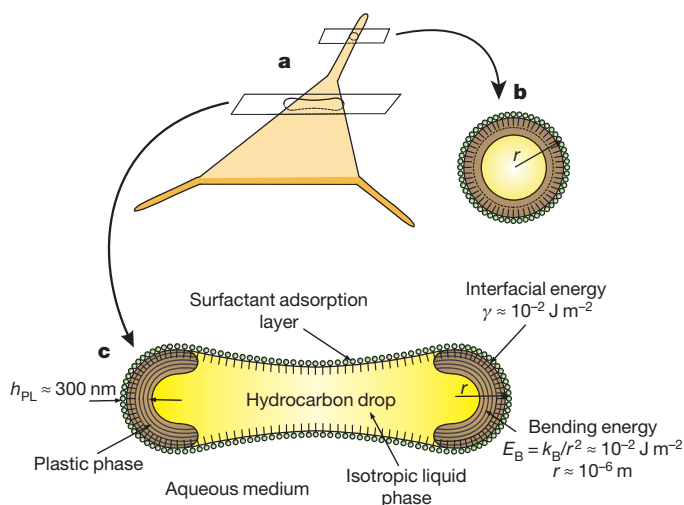




**Figure 3 | Illustrative examples of liquid drop shapes that would be unstable in the absence of elastic properties of the fluid-drop material.** **a**, Image of a triangular platelet with very long cylindrical asperities, protruding from the platelet tips. **b**, Image of rhomboid-shaped liquid drops, some with a hole in their centre, a shape clearly not governed by surface tension only. Such images prove that the fluid drops contain plastic phases, while still containing liquid inside (the oval shape in the puncture determined by surface tension). The coloured, bottom-most shape is the only one frozen. Scale bars, 50  $\mu\text{m}$ .

liquid droplets of various shapes under cross-polarized light showed only faint colours, in contrast to the characteristic intense colours that are typically observed for thick non-isotropic liquid crystal layers<sup>13</sup>. If the interior of the deformed liquid drops were entirely filled with an anisotropic plastic phase, the liquid drops would have appeared more coloured in cross-polarized light. We do observe the appearance of beautiful intense colours, but only in the moment of complete freezing of the droplets, indicating the formation of crystal domains in the frozen particles (Figs 1 and 2).

From these results we concluded that the freezing surfactant adsorption layer induces the formation of a thin layer of a hydrocarbon plastic rotator phase of thickness  $h_{\text{PL}}$ , adjacent to the drop surface, which in turn, drives the observed drop-shape transformations (Fig. 4). Surface-induced formation of liquid crystal phase layers has been observed previously<sup>14,15</sup> and, as shown later, such plastic interfacial sheets possess a sufficiently large bending moment, able to counteract the effects of the hydrocarbon–water interfacial tension that enforces the spherical shape of the common liquid drops. Being submicrometre in thickness, these sheets appear only



**Figure 4 | Schematic presentation of the drop-shape deformation mechanism, driven by the formation of interfacial plastic phases.** **a–c**, Cross-sections of a platelet-shaped particle with protrusions (**a**) are shown in **b**, **c**. In the presence of appropriate surfactant, thin plastic phases (brown regions) with thickness  $h_{\text{PL}}$ , bending elasticity constant  $K_B$  and characteristic curvature of the shaped droplet edges,  $r$ , form at the hydrocarbon–water interface, adjacent to the surfaces with high curvature. The low-temperature-induced, highly curved plastic phases form an energetically favourable expanding frame at the drop edges, which drives the observed shape transformations. For clarity, the figure is not to scale.

with faint colours in cross-polarized light<sup>16</sup>, just as observed in our experiments.

The thickness of these plastic interfacial sheets,  $h_{\text{PL}}$ , could be estimated by considering the balance of the bending and surface area energies of the drop surface. The surface area energy is represented by the interfacial tension,  $\gamma$ , which was measured to be between 5 and 10  $\text{mJ m}^{-2}$  for the surfactant-containing systems studied. The bending energy per unit area of the hydrocarbon–water interface could be estimated<sup>17,18</sup> as  $E_B \approx K_B/r^2$ , where  $K_B$  is the bending elasticity constant and  $r \approx 1 \mu\text{m}$  is the observed characteristic curvature of the shaped droplet edges. Such stable, highly deformed droplets could be formed only if the highly curved phases are energetically favourable, and if  $E_B$  is comparable or bigger than the surface tension  $\gamma$  (both measured in  $\text{J m}^{-2}$ ), which would pull the surface back to the lower spherical curvature. This leads to the minimum estimate for  $K_B \approx 10^{-14} \text{ J}$ , a value much higher than the known<sup>17,18</sup> bending constants of frozen lipid bilayers or surfactant adsorption monolayers,  $K_B \approx 10^{-18} \text{ J}$ . Taking into account<sup>18</sup> that  $K_B$  is proportional to  $h_{\text{PL}}^2$ , we estimate that  $h_{\text{PL}} \approx 300 \text{ nm}$  for the observed deformed drops. Similar values of  $h_{\text{PL}}$  were reported in independent experimental studies<sup>14,15</sup> for surface-induced liquid crystal sheet phases. As already explained, the bending forces resulting from  $h \approx 1.5\text{--}3 \text{ nm}$  for surfactant monolayers and lipid bilayers<sup>17,18</sup>, and the respective bending moment of  $K_B \approx 10^{-18} \text{ J}$ , are far too weak to deform liquid drops for the interfacial tension values measured in our systems.

The elastic layers seem mostly localized at the shape edges and are characterized also by their ‘spontaneous curvature’<sup>17,19</sup>, that is, the curvature that the interface would acquire if no other forces (besides the local intermolecular forces) were involved. All our observations show that the shape transformations are driven by the growth of edges with high spontaneous curvature (small radius of curvature) in one direction, thus forming cylindrical structures in the drop edges. These energetically favourable, cylindrical plastic crystal phases grow in length, thus forming elastic frames that overcome the interfacial tension and stretch the droplets to flatter shapes with high aspect ratios. Indeed, as illustrated in Fig. 3 and Supplementary Video 2, upon puncture, the inside of the rhomboid liquid droplets acquire the minimum circumference dictated by interfacial tension, while preserving intact the outside shaped frame. Figure 1 and Supplementary Video 1 show how the plastic crystals grow and disproportionate into different straight edges to form droplet shapes with longer and longer circumferences, until finally forming highly elongated drops and thin fibres of radius less than  $0.4 \mu\text{m}$ . Such fibres contain all the material in a single ‘edge’ region, probably all composed of a rotator phase and providing an estimate for the spontaneous curvature of these phases.

Our observations with droplets of several linear hydrocarbons show that no specially designed molecules are needed to observe this drop ‘self-shaping’ phenomenon. By selecting hydrocarbons with appropriate chain length, we varied the temperature range for the liquid drop transformations: the tetradecane drops had non-spherical shapes in a range of 0 to 3  $^{\circ}\text{C}$ , hexadecane drops in a range of 9 to 18  $^{\circ}\text{C}$ , and eicosane drops in a range of 30 to 35  $^{\circ}\text{C}$ .

The growth of smectic liquid crystalline fibres in CTAB solutions above the critical micellar concentration has previously been observed<sup>20,21</sup>. However, in these papers only one type of structure was observed (rod-shaped particles) whereas we are able to produce particles with a wide variety of shapes in a controlled manner. The drop-shape sequence of transformations we observe is much richer and probably different in mechanism from those observed previously<sup>20,21</sup>. Also, hexagonal drop shapes of lyotropic liquid crystal phases have been observed before<sup>11</sup>. However, the full array of complex shapes shown in Figs 1 and 2, and the possibility of transforming between them and of capturing them in a frozen state, are novel (see also Extended Data Figs 1 and 2).

Our approach can be used to produce ‘shape-on-demand’ particles, noting that high-aspect-ratio micro/nanoparticles show preferential

internalization in tumour cells<sup>22</sup> and that tissue/organ uptake can be shape specific<sup>23</sup>. A combination with microfluidic techniques<sup>6,24,25</sup> seems particularly suitable to explore the full range of such opportunities and the self-shaping method's governing mechanism of symmetry breaking. By controlling the local temperature profile in the microfluidic channel, custom shaped particle populations or mixtures of specific shapes and various sizes could be produced. The obtained shaped particles and fibres could be used to build hierarchical structures or as sacrificial templates for the production of porous materials with complex morphologies.

We report a novel bottom-up mechanism for morphogenesis and an energy- and material-efficient method for the formation of micro- and nanoscale liquid drops and solid particles with complex shapes. The ability of a single fluid phase to form spontaneously the wide variety of shapes we report could decrease the perceived informational complexity of many structures<sup>26</sup>. This shape-shifting is probably used in nature, it is of clear relevance to the emerging field of active matter, and is expected to be applicable to other rotator-phase- and plastic-phase-forming molecules and biomolecules.

The morphogenesis mechanism we present is expected to stimulate research in a number of fields, as the observed phenomena combine several active research areas, such as capillarity and elasticity, liquid crystal and plastic phases in confined spaces, and surface and bulk nucleation. The process is probably a good platform for investigating phase equilibria, the role of confinement and the melting of two-dimensional crystals, as well as the interplay between liquid crystal defects and surface bending elasticity resulting in shape changes<sup>27,28</sup>. It is of particular interest for elucidating novel mechanisms of symmetry breaking that contribute to understanding the fundamental processes of morphogenesis.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 11 June; accepted 27 October 2015.**

**Published online 9 December 2015.**

1. Turing, A. M. The chemical basis of morphogenesis. *Philos. Trans. R. Soc. Lond.* **237**, 37–72 (1952).
2. Champion, J. A., Katare, Y. K. & Mitragotri, S. Particle shape: a new design parameter for micro- and nanoscale drug delivery carriers. *J. Control. Release* **121**, 3–9 (2007).
3. Xiao, J. & Qi, L. Surfactant-assisted, shape-controlled synthesis of gold nanocrystals. *Nanoscale* **3**, 1383–1396 (2011).
4. Peng, X. *et al.* Shape control of CdSe nanocrystals. *Nature* **2**, 145–150 (2003).
5. Alargova, R. G., Bhatt, K. H., Paunov, V. N. & Velev, O. D. Scalable synthesis of a new class of polymer microrods by a liquid-liquid dispersion technique. *Adv. Mater.* **16**, 1653–1657 (2004).
6. Dendukuri, D. & Doyle, P. S. The synthesis and assembly of polymeric microparticles using microfluidics. *Adv. Mater.* **21**, 4071–4086 (2009).
7. Deitzel, J. M., Kleinmeyer, J., Harris, D. & Beck Tan, N. C. The effect of processing variables on the morphology of electrospun nanofibers and textiles. *Polymer* **42**, 261–272 (2001).
8. Smoukov, S. K. *et al.* Scalable liquid shear-driven synthesis of polymer nanomaterials. *Adv. Mater.* **27**, 2642–2647 (2015).

9. Sirota, E. B. & Herhold, A. B. Transient phase-induced nucleation. *Science* **283**, 529–532 (1999).
10. Ueno, S., Hamada, Y. & Sato, K. Controlling polymorphic crystallization of *n*-alkane crystals in emulsion droplets through interfacial heterogeneous nucleation. *Cryst. Growth Des.* **3**, 935–939 (2003).
11. Jeong, J., Davidson, Z. S., Collings, P. J., Lubensky, T. C. & Yodh, A. G. Chiral symmetry breaking and surface faceting in chromonic liquid crystal droplets with giant elastic anisotropy. *Proc. Natl Acad. Sci. USA* **111**, 1742–1747 (2014).
12. Stone, H. Dynamics of drop deformation and breakup in viscous fluids. *Annu. Rev. Fluid Mech.* **26**, 65–102 (1994).
13. Zhang, L.-Y., Zhang, Q.-K. & Zhang, Y.-D. Design, synthesis, and characterisation of symmetrical bent-core liquid crystalline dimers with diacetylene spacer. *Liq. Cryst.* **40**, 1263–1273 (2013).
14. Wittebrood, M. M., Luijendijk, D. H., Stallinga, S., Rasing, Th. & Musevic, I. Thickness-dependent phase transition in thin nematic films. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **54**, 5232–5234 (1996).
15. Miyano, K. Surface-induced ordering of a liquid crystal in the isotropic phase. *J. Chem. Phys.* **71**, 4108–4111 (1979).
16. Bloss, F. D. *An Introduction to the Methods of Optical Crystallography* (Holt, Rinehart and Winston, 1961).
17. Israelachvili, J. N. *Intermolecular and Surface Forces* (Academic, 2011).
18. Evans, E. A. & Skalak, R. *Mechanics and Thermodynamics of Biomembranes* (CRC, 1980).
19. Jung, H. T., Coldren, B., Zasadzinski, J. A., Iampietro, D. J. & Kaler, E. W. The origins of stability of spontaneous vesicles. *Proc. Natl Acad. Sci. USA* **98**, 1353–1357 (2001).
20. Peddireddy, K. *et al.* Lasing and waveguiding in smectic A liquid crystal optical fibers. *Opt. Express* **21**, 30233–30242 (2013).
21. Peddireddy, K., Kumar, P., Thutupalli, S., Herminghaus, S. & Bahr, C. Myelin structures formed by thermotropic smectic liquid crystals. *Langmuir* **29**, 15682–15688 (2013).
22. Gratton, S. E. *et al.* The effect of particle design on cellular internalization pathways. *Proc. Natl Acad. Sci. USA* **105**, 11613–11618 (2008).
23. Decuzzi, P. *et al.* Size and shape effects in the biodistribution of intravascularly injected particles. *J. Control. Release* **141**, 320–327 (2010).
24. Makgwane, P. R. & Ray, S. S. Synthesis of nanomaterials by continuous-flow microfluidics: a review. *J. Nanosci. Nanotechnol.* **14**, 1338–1363 (2014).
25. Kim, J.-W., Utada, A. S., Fernández-Nieves, A., Hu, Z. & Weitz, D. A. Fabrication of monodisperse gel shells and functional microgels in microfluidic devices. *Angew. Chem. Int. Ed.* **46**, 1819–1822 (2007).
26. Neville, A. C. Molecular and mechanical aspects of helicoid development in plant cell walls. *BioEssays* **3**, 4–8 (1985).
27. Li, Y., Miao, H., Ma, H. & Chen, J. Z. Y. Topological defects of tetratic liquid-crystal order on a soft spherical surface. *Soft Matter* **9**, 11461–11466 (2013).
28. Bowick, M. J. & Giomi, L. Two-dimensional matter: order, curvature and defects. *Adv. Phys.* **58**, 449–563 (2009).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was funded by the European Research Council (ERC) grant to S.K.S., EMATTER (#280078). The study falls under the umbrella of European networks COST MP 1106 and 1305 and the capacity building project BeyondEverest of the European Commission (grant no. 286205).

**Author Contributions** N.D. and S.T. conceived the main idea for the study, designed the experiments and performed most of the result interpretation, N.D. wrote the first draft, I.L. and D.C. performed the experiments and prepared the figures, S.K.S. contributed important ideas for the experiments and data and mechanism interpretation, and edited the text.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.K.S. (sks46@cam.ac.uk).

## METHODS

The alkanes (>99% purity) were purchased from Sigma-Aldrich, and further purified by passing several times through columns of Florisil to remove the polar components. In the absence of surfactant, hydrocarbon–water interfacial tension was measured to be always above  $50 \text{ mN m}^{-1}$  as known for pure alkanes. The hydrocarbon–water interfacial tension (water containing 1.5 wt% surfactant),  $\gamma$ , was measured by drop-shape analysis (instrument DSA100 by Krüss) to be in the range between 5 and  $10 \text{ mN m}^{-1}$  for all surfactant systems studied, in the entire range from room temperature down to the temperatures of drop deformation and freezing.

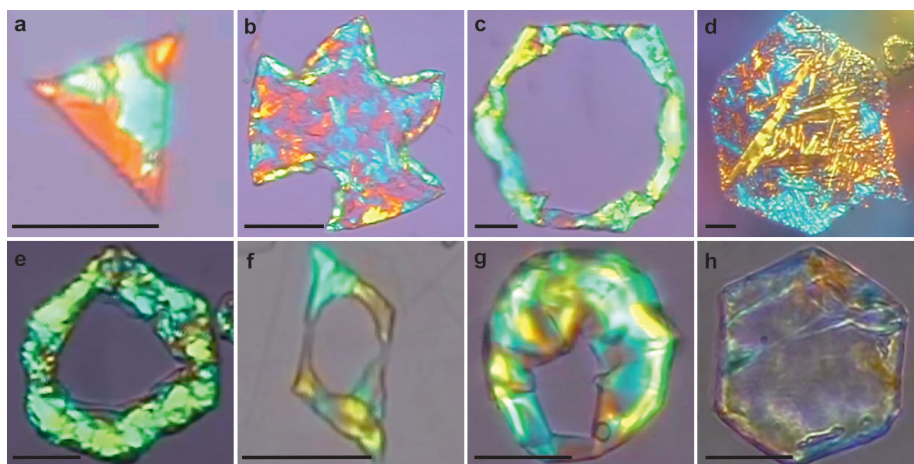
The original hydrocarbon-in-water emulsions were prepared by membrane emulsification with 2, 3, 5 or  $10 \mu\text{m}$  pore-size glass membrane (SPG) in 1.5 wt% solutions of Brij 58 (Fig. 1), Tween 40, Tween 60 or Brij 78 (Fig. 2 and Extended Data Fig. 1), CTAB (Extended Data Fig. 1) or other surfactants (images not shown). All surfactants were chosen to be water soluble with high hydrophilic–lipophilic balance (HLB > 14), so the surfactant would be almost exclusively in the water phase. Extended Data Table 1 summarizes the HLB values.

The emulsion cooling was realized in rectangular glass capillaries with length of 50 mm, width of 1 mm and height of 0.1 mm, enclosed within a custom-made metal cooling chamber, with optical windows for microscope observation (Extended Data Fig. 4). The chamber temperature was controlled by cryo-thermostat (Julabo CF30) and measured close to the emulsion location, using a calibrated thermo-couple probe with an accuracy of  $\pm 0.2^\circ\text{C}$ .

The optical observations were performed with Axioplan and AxioImager M2.m microscopes (Zeiss) in transmitted, cross-polarized white light, with included compensator plate situated after the sample and before the analyser, at  $45^\circ$  with respect to both the analyser and the polarizer. Long-focus objectives  $\times 20$  and  $\times 50$  were used. The drop diameter was determined from microscope images.

The average 'height' of the deformed drops was calculated by dividing the total volume of the drop (calculated from the radius of the initial spherical drops) by the projected area of the non-spherical drop shapes, measured from the microscope images. In this way the aspect ratios, shown in Figs 1 and 2, were determined.

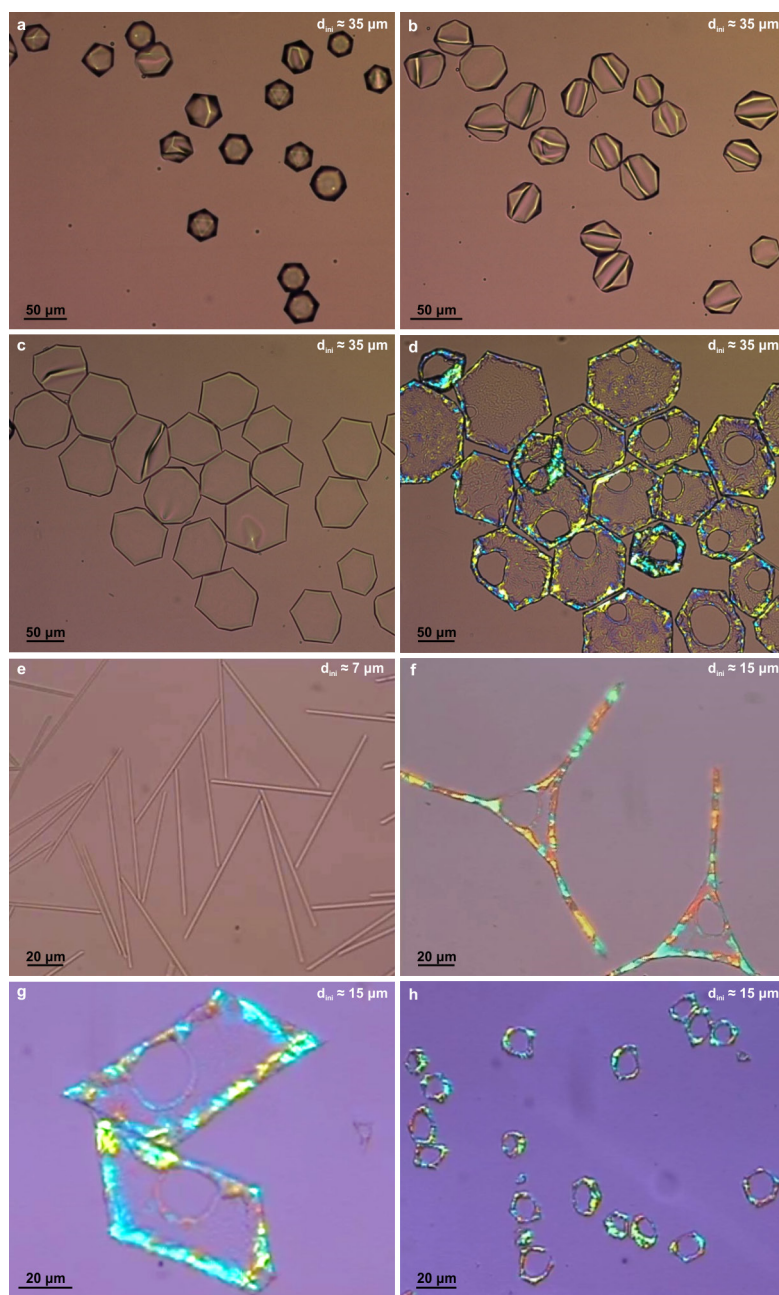




**Extended Data Figure 1 | Solid particles with various shapes.**

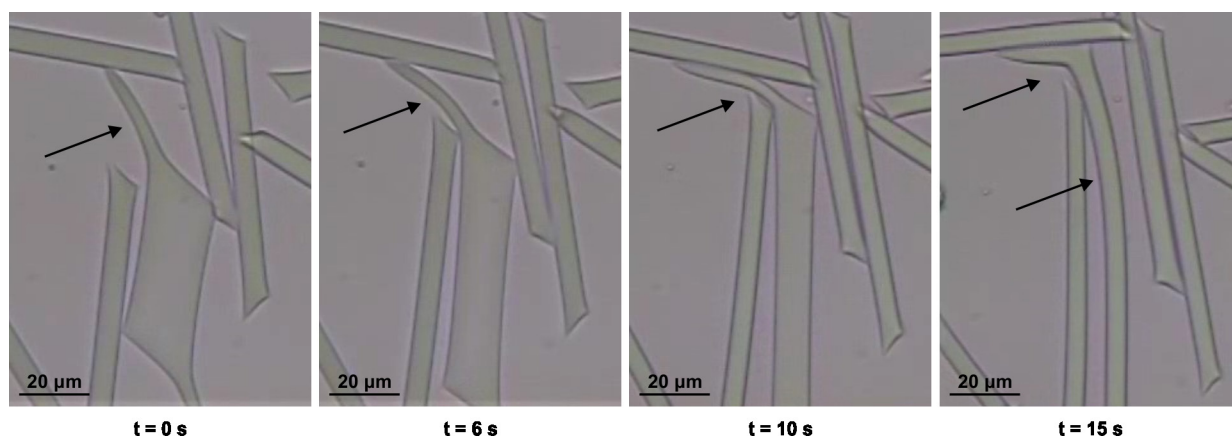
**a–h**, The shapes were obtained by freezing of deformed hexadecane (**a–d**), heptadecane (**e**), tetradecane (**f**), or eicosane (**g, h**) drops in emulsions,

stabilized by 1.5 wt% of different surfactants: **a**, non-ionic Tween 60; **b, c**, non-ionic Brij 78; **d**, cationic CTAB; **e, f**, non-ionic Tween 40; **g, h**, Brij 78. Scale bars, 20  $\mu\text{m}$ .



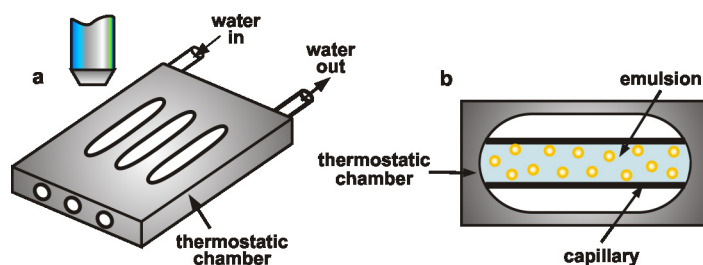
**Extended Data Figure 2 | Snapshot images of multiple hexadecane particles, obtained in 1.5 wt% solutions of various surfactants. a–h, Tween 60 (a–e), Brij 58 (f) and Tween 40 (g, h). a–d, Consecutive images from the evolution of emulsion droplets stabilized by Tween 60.**

**e, Rod-like particles before freezing. f, Frozen triangular particles. g, Frozen tetragonal platelets. h, Frozen toroidal particles.** The initial drop sizes of the particles are indicated on the pictures. Cooling rates are  $0.5 \text{ K min}^{-1}$ , except for **h**,  $2 \text{ K min}^{-1}$ .



**Extended Data Figure 3 | Images proving that the deformed drops are still fluid.** Extending drops collide with each other and bend, as shown with the black arrow. Images of hexadecane drops in 1.5 wt% aqueous solution of Brij 58, cooled with rate of  $1 \text{ K min}^{-1}$ .





**Extended Data Figure 4 | Experimental setup.** **a**, Schematic presentation of the cooling chamber with optical windows, used for microscope observation of the emulsion samples. **b**, The studied emulsions are contained in glass capillaries, placed in the thermostatic chamber and observed through the optical windows.

Extended Data Table 1 | Hydrophilic-lipophilic balance values of the used non-ionic surfactants

Surfactant	HLB value
Brij 58	15.7
Brij 78	15.3
Tween 40	15.5
Tween 60	14.9

HLB, hydrophilic-lipophilic balance.

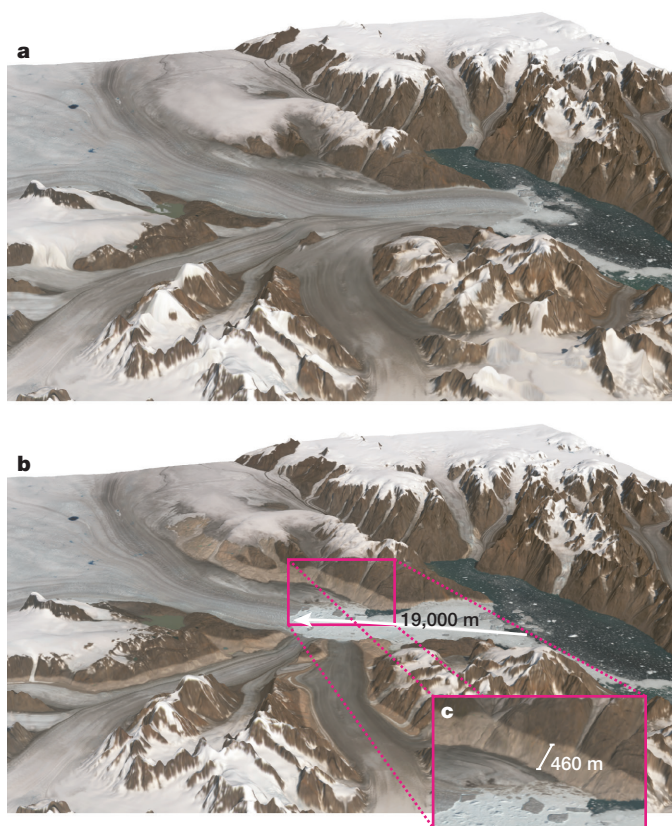
# Spatial and temporal distribution of mass loss from the Greenland Ice Sheet since AD 1900

Kristian K. Kjeldsen<sup>1,2\*</sup>, Niels J. Korsgaard<sup>1\*</sup>, Anders A. Bjørk<sup>1</sup>, Shfaqat A. Khan<sup>3</sup>, Jason E. Box<sup>4</sup>, Svend Funder<sup>1</sup>, Nicolaj K. Larsen<sup>1,5</sup>, Jonathan L. Bamber<sup>6</sup>, William Colgan<sup>4,7</sup>, Michiel van den Broeke<sup>8</sup>, Marie-Louise Siggaard-Andersen<sup>1</sup>, Christopher Nuth<sup>9</sup>, Anders Schomacker<sup>1</sup>, Camilla S. Andresen<sup>4</sup>, Eske Willerslev<sup>1</sup> & Kurt H. Kjær<sup>1</sup>

The response of the Greenland Ice Sheet (GIS) to changes in temperature during the twentieth century remains contentious<sup>1</sup>, largely owing to difficulties in estimating the spatial and temporal distribution of ice mass changes before 1992, when Greenland-wide observations first became available<sup>2</sup>. The only previous estimates of change during the twentieth century are based on empirical modelling<sup>3–5</sup> and energy balance modelling<sup>6,7</sup>. Consequently, no observation-based estimates of the contribution from the GIS to the global-mean sea level budget before 1990 are included in the Fifth Assessment Report of the Intergovernmental Panel on Climate Change<sup>8</sup>. Here we calculate spatial ice mass loss around the entire GIS from 1900 to the present using aerial imagery from the 1980s. This allows accurate high-resolution mapping of geomorphic features related to the maximum extent of the GIS during the Little Ice Age<sup>9</sup> at the end of the nineteenth century. We estimate the total ice mass loss and its spatial distribution for three periods: 1900–1983 ( $75.1 \pm 29.4$  gigatonnes per year), 1983–2003 ( $73.8 \pm 40.5$  gigatonnes per year), and 2003–2010 ( $186.4 \pm 18.9$  gigatonnes per year). Furthermore, using two surface mass balance models<sup>10,11</sup> we partition the mass balance into a term for surface mass balance (that is, total precipitation minus total sublimation minus runoff) and a dynamic term. We find that many areas currently undergoing change are identical to those that experienced considerable thinning throughout the twentieth century. We also reveal that the surface mass balance term shows a considerable decrease since 2003, whereas the dynamic term is constant over the past 110 years. Overall, our observation-based findings show that during the twentieth century the GIS contributed at least  $25.0 \pm 9.4$  millimetres of global-mean sea level rise. Our result will help to close the twentieth-century sea level budget, which remains crucial for evaluating the reliability of models used to predict global sea level rise<sup>1,8</sup>.

We use aerial stereo photogrammetric imagery recorded during the period 1978–1987 to map trimlines and lateral and end moraines associated with the maximum extent of the GIS during the Little Ice Age (LIA<sub>max</sub>), thereby quantifying vertical changes in ice surface elevation between the LIA<sub>max</sub> and 1978–87 (Fig. 1, Methods). To obtain a rate of ice mass loss, the year 1900 AD is assigned as a Greenland-wide time stamp of when the glaciers started to retreat from their LIA<sub>max</sub> position (although we note that this varies regionally and locally<sup>9,12,13</sup>), and 1983 is assigned as the mean year of the aerial observations. Elevation differences after 1983 are derived from airborne and satellite altimetry, combined with a digital elevation model (DEM) developed from the aerial imagery (Methods). We use this geodetic approach to calculate spatially distributed ice thinning patterns and mass balance of the GIS

for three periods (Fig. 2a–c); LIA<sub>max</sub>(1900) to 1983, 1983 to 2003, and 2003 to 2010. We omitted some areas of the GIS because of the lack of LIA data points (Methods).



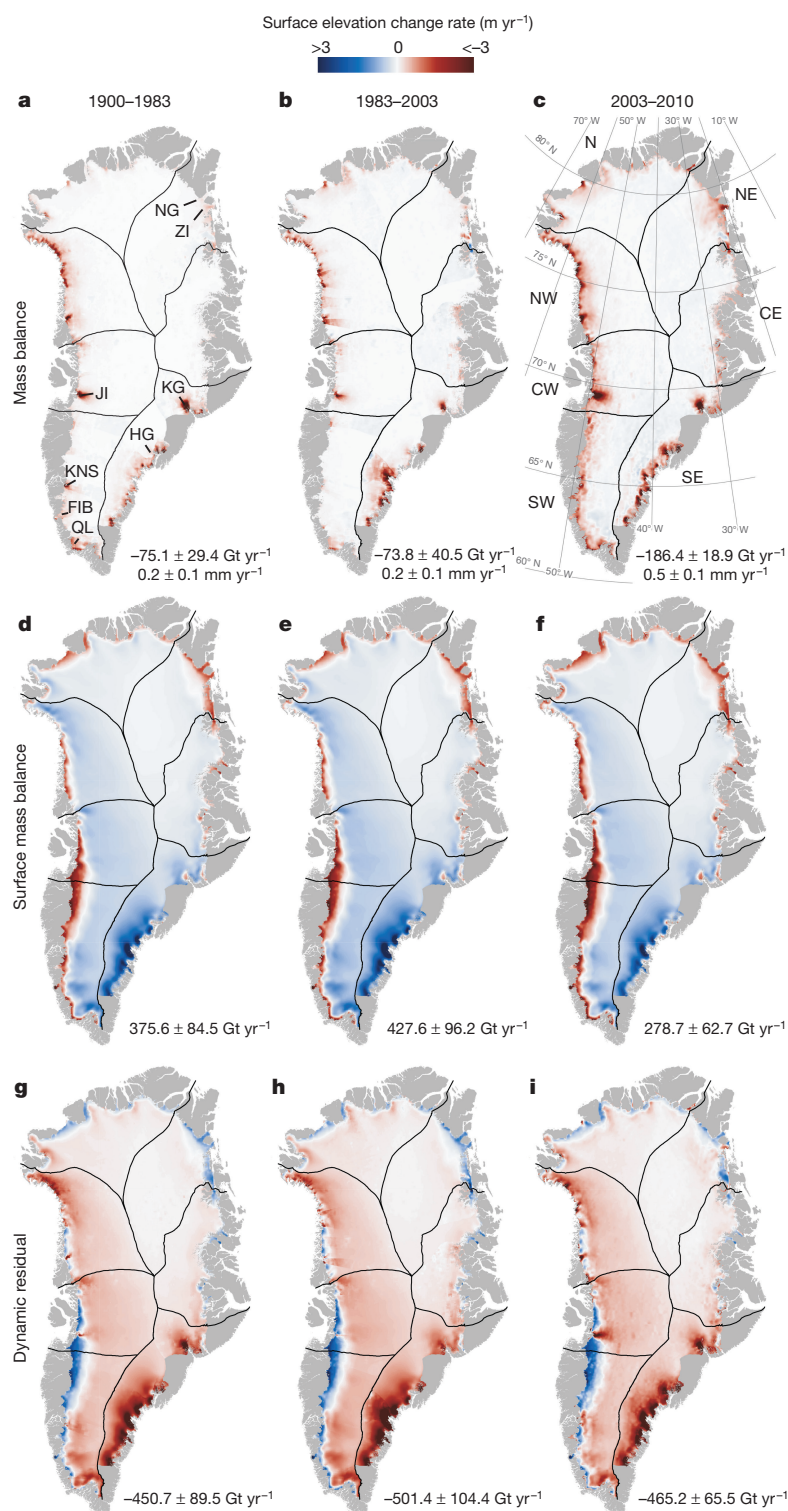
**Figure 1 | Three-dimensional models of Kangerlussuaq Glacier.** **a**, Reconstruction of the LIA<sub>max</sub> ice surface at 1900. **b**, The 2013 ice surface. **c**, Close-up of the northern rim of the 2013 ice surface. The base map is Landsat 8 satellite imagery from 2013. The LIA marks a cold period during which the GIS expanded, often associated with the time interval from 1450–1850<sup>29</sup>. A spectacular indication that the GIS has been shrinking over the last century are the fresh trimlines, that is, the pronounced boundaries between abraded and less abraded bedrock on valley sides and fresh non-vegetated moraines close to the present glacier fronts in many areas of Greenland. Both features are considered to mark the culmination of LIA-glacial advances and to have been mainly formed during the 1700s or at the end of the 1800s<sup>30</sup>.

<sup>1</sup>Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Copenhagen 1350, Denmark. <sup>2</sup>Department of Earth Sciences, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada.

<sup>3</sup>DTU Space—National Space Institute, Technical University of Denmark, Department of Geodesy, Kongens Lyngby 2800, Denmark. <sup>4</sup>Geological Survey of Denmark and Greenland, Department of Marine Geology and Glaciology, Copenhagen 1350, Denmark. <sup>5</sup>Department of Geoscience, Aarhus University, Aarhus 8000, Denmark. <sup>6</sup>Bristol Glaciology Centre, University of Bristol, Bristol BS8 1SS, UK. <sup>7</sup>Department of Earth and Space Science and Engineering, York University, Toronto, Ontario M3J 1P3, Canada. <sup>8</sup>Institute for Marine and Atmospheric Research, Utrecht University, Utrecht 80005, The Netherlands. <sup>9</sup>Department of Geosciences, University of Oslo, Oslo 0316, Norway.

\*These authors contributed equally to this work.





**Figure 2 | Surface elevation change rates in Greenland since the LIA maximum.** The colour scale applies to all panels. **a–c**, Estimates of surface elevation change rates during LIA<sub>max</sub>(1900)–1983 (**a**), 1983–2003 (**b**) and 2003–2010 (**c**). The numbers listed below each panel are the integrated Greenland-wide mass balance estimates expressed as gigatonnes per year and as millimetre per year GMSL equivalents. The associated uncertainties include an uncertainty related to the scaling approach, an error related to observed changes during 2003–2010, and an uncertainty related to the scaling of the point-based observations. **d–f**, Total estimates of surface elevation change rates due to SMB fluctuations, using revised SMB

estimates from ref. 10 during LIA<sub>max</sub>(1900)–1983 (**d**), 1983–2003 (**e**), and 2003–2010 (**f**). **g–i**, The dynamically driven residual in elevation change rates during LIA<sub>max</sub>(1900)–1983 (**g**), 1983–2003 (**h**), and 2003–2010 (**i**). Negative values indicate mass loss. Uncertainties are reported as  $1\sigma$ . Labels in **a** refer to Jakobshavn Isbræ (JI), Kangerlussuaq Glacier (KG), Helheim Glacier (HG), Zachariae Isstrøm (ZI), and Nioghalvfjordsfjorden Glacier (NG), respectively. Labels in **c** refer to north (N), northeast (NE), central east (CE), central west (CW), northwest (NW), southwest (SW) and southeast (SE), respectively.

**Table 1 | Mass balance and components LIA<sub>max</sub>(1900)–2010**

		GIS	SW	CW	NW	N	NE	CE	SE
LIA <sub>max</sub> (1900)–1983	Mass balance	−75.1 ± 29.4	−8.7 ± 4.4	−7.9 ± 4.1	−27.6 ± 6.2	−2.9 ± 3.7	2.8 ± 4.1	−0.1 ± 2.3	−30.6 ± 4.5
	Revised SMB estimates <sup>10</sup>	375.6 ± 84.5	39.9 ± 9.0	55.3 ± 12.4	65.6 ± 14.8	20.2 ± 4.5	8.5 ± 1.9	21.2 ± 4.8	164.9 ± 37.1
	Dynamic residual	−450.7 ± 89.5	−48.6 ± 10.0	−63.2 ± 13.1	−93.2 ± 16.0	−23.1 ± 5.9	−5.7 ± 4.5	−21.4 ± 5.3	−195.5 ± 37.4
1983–2003	Mass balance	−73.8 ± 40.5	−3.0 ± 2.9	−5.9 ± 3.4	−23.4 ± 6.4	−4.7 ± 6.8	0.7 ± 8.9	0.6 ± 6.5	−38.0 ± 5.5
	Revised SMB estimates <sup>10</sup>	427.6 ± 96.2	50.1 ± 11.3	63.3 ± 14.2	69.2 ± 15.6	24.0 ± 5.4	9.3 ± 2.1	25.7 ± 5.8	186.0 ± 41.9
	Dynamic residual	−501.4 ± 104.4	−53.2 ± 11.6	−69.2 ± 14.6	−92.6 ± 16.8	−28.7 ± 8.7	−8.6 ± 9.2	−25.2 ± 8.7	−224.0 ± 42.2
2003–2010	Mass balance	−186.4 ± 18.9	−29.7 ± 4.6	−28.6 ± 3.0	−47.4 ± 2.1	−15.6 ± 1.3	−7.2 ± 2.0	−7.4 ± 2.2	−50.5 ± 3.6
	Revised SMB estimates <sup>10</sup>	278.7 ± 62.7	6.9 ± 1.6	48.2 ± 10.9	50.9 ± 11.5	6.0 ± 1.4	5.1 ± 1.2	18.0 ± 4.0	143.5 ± 32.3
	Dynamic residual	−465.2 ± 65.5	−36.6 ± 4.9	−76.8 ± 11.3	−98.4 ± 11.7	−21.7 ± 1.9	−12.3 ± 2.4	−25.4 ± 4.6	−194.0 ± 32.5

Estimates of mass balance derived using the geodetic approach, the revised SMB estimates from ref. 10, and the dynamic residual of the GIS and the individual regions. Units, Gt yr<sup>−1</sup>.

Figure 2a–c illustrates the annual mass balance for the three periods. We calculate a net mass loss of  $6,233 \pm 2,436$  Gt ( $75.1 \pm 29.4$  Gt yr<sup>−1</sup>) between the onset of glacial retreat from the LIA<sub>max</sub> position (which we take to be 1900, as defined above) and 1983 (Fig. 2a). In northwest Greenland, where the majority of the ice sheet discharges through marine outlet glaciers, we find substantial and widely distributed thinning, leading to a mass loss of  $27.6 \pm 6.2$  Gt yr<sup>−1</sup>, corresponding to 37% of the total mass loss (Table 1). In west and southwest Greenland, we find peripheral thinning concentrated near the two large marine outlet glaciers Jakobshavn Isbræ and Kangerlussuaq Nunata Sermia. Substantial changes also occurred at the land-based glaciers Frederikshåb Isblink and Qassimiut Lobe, the latter being intersected by relatively small fjords draining its eastern part. Along the southeast coast, a region dominated by large marine outlet glaciers, thinning was extensive, in some areas propagating almost to the ice divide, causing a mass loss of  $30.6 \pm 4.5$  Gt yr<sup>−1</sup> (41% of the total). Here two of the largest outlet glaciers in Greenland<sup>3</sup>, Kangerlussuaq Glacier and Helheim Glacier, show distinctly different patterns, with Kangerlussuaq Glacier being the single largest point source of mass loss ( $10.6 \pm 1.2$  Gt yr<sup>−1</sup>), accounting for 14% of the total ice sheet mass loss during this period, while Helheim Glacier appears to have been near balance (mass gain equivalent to mass loss), despite the fact that front positions reveal a considerable inter-period variability<sup>9,14</sup> of about 9 km. In east, northeast, and north Greenland thinning is less extensive and in some areas the ice margin remains at or very close to its LIA<sub>max</sub> position, which in northern Greenland may be attributed to the confining effect of semi-permanent fjord ice on ice discharge<sup>15</sup>. The inference of persistent mass loss of the GIS since LIA<sub>max</sub> may challenge the assumption of a near-balance ice sheet during the 1961–1990 period that is generally invoked to partition recent mass loss (that is, determine mass loss either by surface processes or ice discharge), and thus a failure to acknowledge mass loss during the reference period can result in overestimating the recent ice mass lost owing to surface mass balance (SMB) and ice dynamic processes<sup>16</sup>.

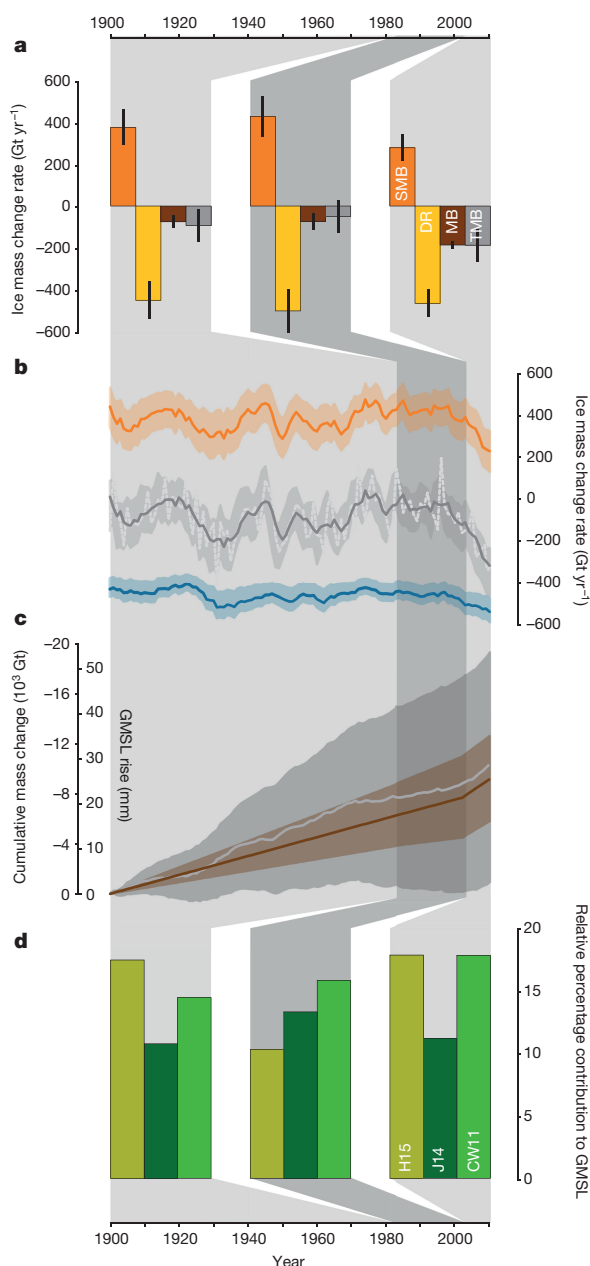
We calculate a total mass loss of  $1,475 \pm 809$  Gt ( $73.8 \pm 40.5$  Gt yr<sup>−1</sup>) for the period 1983–2003 (Fig. 2b). In general, peripheral ice thinning was less widespread and many of the largest outlet glaciers showed a decreasing mass loss (Table 1). During this period, 83% of the total mass loss occurred in the northwest and southeast while Jakobshavn Isbræ alone accounted for 6%, indicating that loss in the remainder of the ice sheet was limited. Interestingly, a comparison of our estimate with studies that have higher temporal resolution suggests that most of the overall, ice-sheet-wide mass loss that we record during 1983–2003 occurred in the late 1990s and early 2000s<sup>17</sup> following a more stable period in the 1980s<sup>3</sup>.

Between 2003 and 2010, we estimate a mass loss of  $1,305 \pm 132$  Gt ( $186.4 \pm 18.9$  Gt yr<sup>−1</sup>), based on the ice mask we employed (Fig. 2c); when we used the same ice mask as ref. 18 (Methods, Extended

Data Fig. 3) we obtain a mass loss of  $250.1 \pm 21.2$  Gt yr<sup>−1</sup>, which is comparable to other studies<sup>2,17</sup>. We find that 2003–2010 mass loss not only more than doubled relative to the 1983–2003 period, but also relative to the net mass loss rate throughout the twentieth century. This latter observation corroborates other studies which have inferred accelerated mass loss in the early twenty-first century relative to the late twentieth century<sup>3,5,19</sup>. Many areas currently undergoing changes are identical to those which underwent considerable thinning throughout the twentieth century, with the exception of Helheim Glacier and the Nioghalvfjærdssjorden Glacier (Fig. 2a–c). Consequently, comparing the twentieth-century thinning pattern to that of the last decade, and assuming a similar warming pattern, we suggest that the overall present mass loss pattern will persist for mass loss in the near future, at least until major marine outlet glaciers become land-terminating; though this may be biased because recent observations from northeast Greenland suggest a considerable acceleration in mass loss from Nioghalvfjærdssjorden Glacier, following at least 20 years of dormancy, and from the Zachariae Isstrøm glacier<sup>18</sup>.

To assess the SMB and ice dynamic components of the twentieth-century mass balance we use updated SMB estimates from ref. 10 (Fig. 2d–f), which have been refined by implementing a more physically based meltwater retention scheme, and calibrating for better agreement with RACMO2.1/GR<sup>11</sup> during the period 1960–2012 (Methods). The ice dynamic residual is calculated by subtracting surface lowering caused by SMB processes from the reconstructed total mass balance (Fig. 2g–i) and is largely similar to the SMB pattern, though with positive values in the ablation zone and negative values in the accumulation zone. This general pattern is suggestive of an ice sheet close to balance; however, the residual also includes elevation trends due to forcing that is not included in the SMB model we employ. Perhaps unsurprisingly, we find a large dynamic contribution to the mass balance in the southeast and northwest, both dominated by marine-terminating glaciers, whereas in other regions the land-terminating ice sheet margin exhibits a positive dynamic mass contribution to compensate for the lowering of the ice surface due to SMB processes. Our results suggest that variability of the dynamic term of the GIS mass balance during the three intervals, which are LIA<sub>max</sub>(1900)–1983, 1983–2003 and 2003–2010, is less than its associated uncertainties (Fig. 3a). Previous results have attributed the mass loss in 2000–2008 equally to decreasing SMB and to increasing discharge<sup>20</sup>, while estimates for more recent periods suggest that decreasing SMB is becoming the dominant driver for increasing mass loss<sup>18,21</sup>. Here we find that although short-term dynamic variability may affect the mass balance<sup>18,21–23</sup>, on a centennial timescale the dominant driver for changes in the GIS mass balance so far appears to be variability in SMB (Fig. 3a).

The temporal variability of the mass balance during the twentieth century is computed as the difference between the updated SMB



**Figure 3 | Mass balance and implication of GMSL.** **a**, Revised estimates of SMB from ref. 10 (orange bars), the ice dynamic residual (DR, yellow bars), mass balance based on the geodetic method (MB, dark brown bars), and mass balance based on the temporal mass balance approach (grey bars) covering the three periods LIA<sub>max</sub>(1900)–1983, 1983–2003 and 2003–2010. Black lines represent the associated 1σ uncertainty ranges. The results suggest that variability in SMB affects long-term mass loss more strongly than does dynamic variability, which on a centennial timescale is more constant. **b**, The orange trace shows the 5-year running mean of the revised SMB estimates from ref. 10, the blue line represents the ice discharge modelled as a function of runoff using a 6-year trailing mean, and the dotted grey and solid grey lines show the yearly and 5-year running mean mass balance, respectively. The shaded areas reflect the associated 1σ uncertainty range (Methods). **c**, Cumulative mass change since LIA<sub>max</sub>(1900) from the geodetic approach (brown line) and from the temporal mass balance reconstruction (grey line), and the shading gives the 1σ uncertainty ranges. **d**, The bars show the contribution of mass loss of the GIS relative to different solutions of the twentieth century GMSL rise from ref. 26 (H15, light green), ref. 27 (J14, dark green), and ref. 28 (CW11, green). Our result shows the minimum relative input of the GIS to sea level rise, which ranges between 10% and 18% during LIA<sub>max</sub>(1900)–2010, supporting a substantial contribution from Greenland during the twentieth century.

estimates of ref. 10 and modelled ice discharge derived as a function of runoff<sup>5,24</sup>, using a 6-year trailing mean, and ice discharge data from ref. 21 (Methods). During the period LIA<sub>max</sub>(1900)–2010 we find a mass loss of  $10,071 \pm 8,580$  Gt, which, despite the use of a smaller ice mask, is slightly higher than that of ref. 5. Although this ancillary temporal mass balance method is particularly sensitive to the ice discharge proxy employed, we find good absolute agreement with the mass loss of  $9,013 \pm 3,378$  Gt found using the geodetic method presented above; this adds constraints and confidence to the results presented here.

Our temporal mass balance method suggests considerable variability in the mass balance during the twentieth century (Fig. 3b). The greatest negative mass balance rates occurred during the late 1920s and early 1930s, a period during which the rate of air temperature increase was higher than during the past decade<sup>14,25</sup>, and which also coincides with extensive glacier retreat in southeast Greenland<sup>14</sup>. Following substantially lower or even nearly zero negative mass balance rates during the 1940s, our model results suggests mass loss rates during the 1950s and 1960s that are similar to those observed during the late 1990s and the early twenty-first century<sup>17</sup>. In the period covering the 1960s to the 1980s our results are comparable to other modelling results that generally suggest net mass loss during the 1960s and an ice sheet near balance during the 1970s to 1980s (ref. 3).

In the Fifth Assessment Report of the Intergovernmental Panel on Climate Change<sup>8</sup>, the twentieth-century global-mean sea level (GMSL) budget was assessed by comparing estimates derived from tide-gauges against observations of the different contributors, leading to unassigned residual sea level rise during 1901–1990. However, in ref. 8, no observational records of the contribution from GIS or the Antarctic Ice Sheet before 1993 are included. The failure to close the GMSL budget for the period 1901–1990 has been attributed to underestimation of the individual contributor factors, including the polar ice sheets<sup>1,8</sup>. A recent study recalculated the twentieth-century GMSL using a probabilistic technique only to find a considerably lower rate of twentieth-century GMSL rise before 1993, thus closing the budget without including contributions from the polar ice sheets<sup>26</sup>. However, our results show that during the twentieth century the GIS contributed substantially to GMSL rise (Fig. 3c).

In particular, the geodetic approach that is based on observations from aerial imagery, which indicates considerable thinning along the margin of the ice sheet, is regarded as a conservative minimum estimate of mass loss (Methods). We find using the geodetic approach a total mass loss of  $9,013 \pm 3,378$  Gt from LIA<sub>max</sub>(1900) to 2010, equivalent to  $25.0 \pm 9.4$  mm of GMSL rise, and a mass loss of  $10,071 \pm 8,580$  Gt (equivalent to  $28.0 \pm 23.8$  mm GMSL rise) using our temporal mass balance method, and thus our results suggest that the GIS has contributed significantly to the twentieth-century sea level budget. Combining our geodetic-based results with recent GMSL reconstructions<sup>26–28</sup> shows that in 1900–1983 the contribution from the GIS to GMSL rise ranged between 11% and 17%; in 1983–2003 it ranged between 10% and 16% and in 2003–2010 it ranged between 11% and 18% (Fig. 3d). Using the same ice mask as ref. 18 we find that during 2003–2010 the contribution to sea level rise ranged between 15% and 24%.

Thus far, any attempt to reconstruct long-term surface elevations beyond the scope of individual outlet glaciers has been prevented by the lack of a suitable Greenland-wide elevation model that would allow accurate observations of moraine and trimline heights representing the maximum ice sheet extent during the LIA. Our study provides 110 years of spatial and temporal mass balance of the GIS and in addition centennial estimates of the SMB and dynamic terms of the mass balance. Finally, our conservative, observation-based results, showing considerable mass loss during the twentieth century from the GIS, minimize the unassigned residual GMSL rise during 1901–1990. This will help to close the twentieth-century GMSL budget, which is crucial for evaluating the reliability of modelling contributions to past sea level rise, and hence for increasing confidence in projections of sea level rise<sup>1,8</sup>.



**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 4 May; accepted 26 October 2015.**

- Gregory, J. M. *et al.* Twentieth-century global-mean sea level rise: is the whole greater than the sum of the parts? *J. Clim.* **26**, 4476–4499 (2013).
- Khan, S. A. *et al.* Greenland ice sheet mass balance: a review. *Prog. Phys.* **78**, 046801 (2015).
- Rignot, E., Box, J. E., Burgess, E. & Hanna, E. Mass balance of the Greenland ice sheet from 1958 to 2007. *Geophys. Res. Lett.* **35**, L20502 (2008).
- Yanga, L. A southern Greenland ice sheet glacier discharge reconstruction: 1958–2007. *Phys. Procedia* **22**, 292–298 (2011).
- Box, J. E. & Colgan, W. Greenland ice sheet mass balance reconstruction. Part III: marine ice loss and total mass balance (1840–2010). *J. Clim.* **26**, 6990–7002 (2013).
- van de Wal, R. & Oerlemans, J. An energy balance model for the Greenland ice sheet. *Global Planet. Change* **9**, 115–131 (1994).
- Zuo, Z. & Oerlemans, J. Contribution of glacier melt to sea-level rise since AD 1865: a regionally differentiated calculation. *Clim. Dyn.* **13**, 835–845 (1997).
- Church, J. A. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) 1137–1216 (Cambridge Univ. Press, 2013).
- Khan, S. A. *et al.* Glacier dynamics at Helheim and Kangerdlugssuaq glaciers, southeast Greenland, since the Little Ice Age. *Cryosphere* **8**, 1497–1507 (2014).
- Box, J. E. Greenland ice sheet mass balance reconstruction. Part II: surface mass balance (1840–2010)\*. *J. Clim.* **26**, 6974–6989 (2013).
- van Angelen, J. H., van den Broeke, M. R. & van de Berg, W. J. Momentum budget of the atmospheric boundary layer over the Greenland ice sheet and its surrounding seas. *J. Geophys. Res.* **116**, D10101 (2011).
- Csatho, B. M., Schenk, T., van der Veen, C. J. & Krabill, W. B. Intermittent thinning of Jakobshavn Isbræ, West Greenland, since the Little Ice Age. *J. Glaciol.* **54**, 131–144 (2008).
- Lea, J. M. *et al.* Terminus-driven retreat of a major southwest Greenland tidewater glacier during the early 19th century: Insights from glacier reconstructions and numerical modelling. *J. Glaciol.* **60**, 333–344 (2014).
- Björk, A. A. *et al.* An aerial view of 80 years of climate-related glacier fluctuations in southeast Greenland. *Nat. Geosci.* **5**, 427–432 (2012).
- Higgins, A. K. North Greenland glacier velocities and calf ice production. *Polarforschung* **60**, 1–23 (1990).
- Colgan, W. *et al.* Greenland high-elevation mass balance: inference and implication of reference period (1961–90) imbalance. *Ann. Glaciol.* **56**, 105–117 (2015).
- Shepherd, A. *et al.* A reconciled estimate of ice sheet mass balance. *Science* **338**, 1183–1189 (2012).
- Khan, S. A. *et al.* Sustained mass loss of the northeast Greenland ice sheet triggered by regional warming. *Nature Clim. Change* **4**, 292–299 (2014).
- Rignot, E., Velicogna, I., van den Broeke, M. R., Monaghan, A. & Lenaerts, J. T. M. Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level rise. *Geophys. Res. Lett.* **38**, L05503 (2011).
- van den Broeke, M. *et al.* Partitioning recent Greenland mass loss. *Science* **326**, 984–986 (2009).
- Enderlin, E. M. *et al.* An improved mass budget for the Greenland ice sheet. *Geophys. Res. Lett.* **41**, 866–872 (2014).
- Kjær, K. H. *et al.* Aerial photographs reveal late-20th-century dynamic ice loss in northwestern Greenland. *Science* **337**, 569–573 (2012).
- Howat, I. M., Joughin, I. R. & Scambos, T. A. Rapid changes in ice discharge from Greenland outlet glaciers. *Science* **315**, 1559–1561 (2007).
- Bamber, J., van den Broeke, M. R., Ettema, J., Lenaerts, J. & Rignot, E. Recent large increases in freshwater fluxes from Greenland into the North Atlantic. *Geophys. Res. Lett.* **39**, L19501 (2012).
- Box, J. E., Yang, L., Bromwich, D. H. & Bai, L.-S. Greenland Ice Sheet surface air temperature variability: 1840–2007\*. *J. Clim.* **22**, 4029–4049 (2009).
- Hay, C. C., Morrow, E., Kopp, R. E. & Mitrovica, J. X. Probabilistic reanalysis of twentieth-century sea-level rise. *Nature* **517**, 481–484 (2015).
- Jevrejeva, S., Moore, J. C., Grinsted, A., Matthews, P. & Spada, G. Trends and acceleration in global and regional sea levels since 1807. *Global Planet. Change* **113**, 11–22 (2014).
- Church, J. A. & White, N. J. Sea-level rise from the late 19th to the early 21st century. *Surv. Geophys.* **32**, 585–602 (2011).
- Kobashi, T. *et al.* On the origin of multidecadal to centennial Greenland temperature anomalies over the past 800 yr. *Clim. Past* **9**, 583–596 (2013).
- Weidick, A., Bennike, O., Citterio, M. & Nørgaard-Pedersen, N. Neoglacial and historical glacier changes around Kangarsuneq fjord in southern West Greenland. *Geol. Surv. Denmark Greenland Bull.* **27**, 1–68, <http://www.geus.dk/publications/bull/nr27/index-uk.htm> (2012).

**Acknowledgements** This study would not have been possible without the aid of The Danish Geodata Agency (GST), who gave us access to their historical aerial photographs. This work is a part of the X\_Centuries project funded by the Danish Council for Independent Research (FNU) (grant number DFF-0602-02526B) and the Centre for GeoGenetics supported by the Danish National Research Foundation (DNRF94). K.K.K. acknowledges support from the Danish Council for Independent Research (FNU) and the Sapere Aude: DFF-Research Talent programme (grant number DFF-4090-00151). J.E.B., K.H.K. and N.K.L. acknowledge support by the GeoCenter Denmark (“Multi-millennial ice volume changes of the Greenland ice sheet”). S.A.K. acknowledges supports from the Carlsberg Foundation (grant number CF14-0145) and the Danish Council for Independent Research (FNU) (grant number DFF-4181-00126). M.v.d.B. acknowledges support from the Netherlands Polar Program of the Netherlands Organization of Scientific Research (NWO). C.N. acknowledges support by the European Research Council (EUFP7/ERC grant number 320816). We thank A. J. Long, S. A. Woodroffe, B. M. Vinther, and R. Hukmans for contribution during the early phase of this study.

**Author Contributions** K.K.K. and K.H.K. designed and conducted the study. N.J.K. did photogrammetric modelling and aero-photogrammetric DEM processing, and quality control and validation with C.N. K.K.K. undertook the Geographical Information System analysis. A.A.B. conducted the manual photogrammetry measurements. S.A.K. carried out analysis of surface elevation data, developed the scaling method, and made the mass balance calculations. J.E.B., J.L.B., and M.v.d.B. provided SMB model and context. W.C., J.E.B., and K.K.K. performed temporal discharge and mass balance modelling. S.F. and N.K.L. provided the historical context of ice sheet extent. All authors contributed to discussion and writing of the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.H.K. ([kurtk@snm.ku.dk](mailto:kurtk@snm.ku.dk)).

## METHODS

Elevation changes between  $LIA_{\max}$  and 1978–87 are derived from direct observations of  $LIA_{\max}$  moraines and trimlines and the ice surface in vertical stereo photogrammetric imagery recorded during 1978–87. Changes are extrapolated to the ice sheet interior using a scale-value approach based on aerial and satellite altimeter data from the period 2003–2010, with site-specific interpolations in 82 basins around the ice sheet. The same scaling approach is used to derive changes from 1978–87 to 2003. For the entire  $LIA_{\max}$ –2010 period we use annual estimates from a SMB model<sup>10</sup> to quantify mass balance processes and to assess temporal variability of the mass balance components through time. In-depth descriptions of the methods used are provided below.

**Geometric approach to derive surface elevation changes.** Previously, mass balance estimates of the entire GIS have been based on modelling efforts that rely on empirical relations between SMB and ice discharge<sup>3–5</sup> or energy balance modelling<sup>6,7</sup>. Geometric approaches have been applied to Jakobshavn Isbræ<sup>12</sup>, outlet glaciers in Patagonia<sup>31</sup>, and land-terminating glaciers on James Ross Island, Antarctic<sup>32</sup>, by mapping trimlines and lateral and end moraines. These studies, however, focus on single-outlet glaciers from the GIS, smaller ice caps, or small isolated glaciers, respectively. Each study used varying methods to account for elevation changes at higher elevations inland, for example, finding upper boundary changes by vertical shifting of the contemporary equilibrium line altitude based on lapse rate temperature reconstructions<sup>31</sup>, or by the vertical difference between trimlines and ice surfaces to provide elevation offsets and thereby estimate mass loss<sup>32</sup>.

Here, we outline the geometric method we deployed that allows us to translate point observations of former ice margin position to an ice-sheet-wide mass balance. The height of an equilibrium glacier or ice sheet profile can be expressed as (see, for example, ref. 33):

$$h(x) = H \left( 1 - \left( \frac{x}{L} \right)^{1+1/n} \right)^{1/(2+2/n)} \quad (1)$$

where  $H$  is the surface elevation at the ice divide,  $L$  is the length of the ice sheet profile,  $h$  is the surface elevation at distance  $x$  from the ice divide, and the exponent  $n$  is a constant. This relation assumes no sliding, a flat bed, uniform accumulation, and constant flowband width. Here, we apply it to show that by using elevation changes between time  $t_1$  and time  $t_2$  it is possible to estimate elevation changes during another period, for example, time  $t_1$  and time  $t_3$ , or to extend the estimate further, for example between time  $t_3$  and time  $t_4$ , by scaling elevation changes of the known period. Subsequently, the approach is assessed using observations at three main outlet glaciers in Greenland: Kangerlugssuaq Glacier, Helheim Glacier, and Jakobshavn Isbræ.

However, first we consider three ice surface profiles  $h$ , from the ice divide to the ice margin, using typical values for the GIS. Each surface elevation profile represents one of the time steps  $t_1$ ,  $t_2$  and  $t_3$ . The glacier length from the ice divide  $L$  is in this example  $x = 200,000$  m at  $t_1$ , and changes by 1,000 m for each time step, while the ice divide height  $H$  is kept constant at 3,300 m, and the exponent  $n$  is set to 3 (Extended Data Fig. 1a). We simulate surface changes by changing the glacier length  $L$  (this corresponds to advance or retreat of the ice margin), and thus surface elevation changes at  $x$  are governed by the total length of the ice profile. In our example, the ice retreats from the initial time step  $t_1$  by 1,000 m and the ice surface lowers at  $t_2$ . Next, we predict the ice profile at  $t_3$  by applying a scale-value and define the predicted profile  $3$  ( $h_{\text{pre}_t3}$ ) (Extended Data Fig. 1b) as:

$$h_{\text{pre}_t3} = h_{t1} + S(h_{t2} - h_{t1}) \quad (2)$$

where  $S$  is a constant.

Comparing the elevation changes between  $h_{t1}$  and  $h_{t3}$  ( $dh_{t1t3}$ ) and those between  $h_{t1}$  and  $h_{\text{pre}_t3}$  ( $dh_{t1t3\_pre}$ ), derived using equation (2) and an  $S$ -value of 2.2, shows overall agreement, though also differences near the margin (Extended Data Fig. 1c). However, here the surface profile  $h_{t1}$  is part of both the input and of the output. To generate a predicted difference where the same timestamp (for example,  $t_1$ ) is not incorporated in the input and the output, the  $S$  can be altered and an 'independent'  $dh$  estimate may be calculated. Here,  $h_{t2} - h_{t1}$  ( $dh_{t1t2}$ ) and an  $S$ -value of 1.2 simulates  $dh_{t3t4\_pre}$ , which shows overall agreement with  $h_{t3} - h_{t4}$  ( $dh_{t3t4}$ ), but again also differences near the margin (Extended Data Fig. 1d). Nevertheless, it implies that if  $dh_{t1t2}$  and  $dh_{t3t4}$  are both known the constant  $S$  can be derived as the ratio between these values.

Extended Data Fig. 1e shows the difference between profile  $h_{t3}$  and  $h_{\text{pre}_t3}$  using a constant  $S$ . Over large parts of the profile the difference is small ( $< 1$  m), however, near the margin differences increase to tens of metres. We use the difference as an expression of the constant  $S$  and denote it  $\sigma_{\text{smeth}}$  (and include it in our

mass balance uncertainty calculations). Extended Data Fig. 1f shows change in elevation (in metres) between two timestamps as a function of surface elevation. Thinning is largest at lower elevations, but drops rapidly and become close to 0 at  $h > 2,500$  m.

We note that, considering the differences near the margin (Extended Data Fig. 1e), the profile approach employed does not work (well) near the terminus of marine-terminating outlet glaciers. As discussed in Methods section 'Uncertainties and conservative mass balance estimates', however, we use an ice mask derived from aerial images recorded during 1978–87, and thus the large differences between simulated and predicted surface profiles, that is,  $\sigma_{\text{smeth}}$  (Extended Data Fig. 1e) and large elevation changes (Extended Data Fig. 1f) at low elevations are not included in the estimate of the period between  $LIA_{\max}$  (1900) to 1978–87.

The approach presented here is founded in the relation in equation (1), for which certain assumptions are made. These assumptions are violated for a large part of the ice sheet. For instance, basal sliding is considerable near marine-terminating outlet glaciers, which combined drain 88% of the ice sheet, and over the majority of the entire ice sheet basal-sliding motion dominates over internal deformation<sup>34</sup>. Extended Data Fig. 2 provides three examples where we apply our approach to major marine-terminating outlet glaciers. Here, we compare elevation changes derived using our scaling approach with elevation changes derived from the DEM (see Methods section 'Photogrammetric DEM 1978–87') and 2003 NASA Airborne Topographic Mapper (ATM) flight lines<sup>35</sup>. We find good agreement (within uncertainties) between the observed and predicted elevation change rates. The examples in Extended Data Fig. 2 illustrate the validity of our approach in fast-flowing areas, where basal sliding is considerable, the bed is not flat, the accumulation is non-uniform, and the width of the flowband is not constant, that is, where the assumptions of the relation in equation (1) are violated. Moreover, the combined uncertainty that we estimate includes an uncertainty related to the scaling approach ( $\sigma_{\text{smeth}}$ ), an error related to changes during 2003–2010 ( $dh_{\text{solid}}$ ) (see Methods section '2003–2010 elevation changes from air- and space-borne laser altimetry'), and an uncertainty related to the scaling of point-based observations, for example,  $dh_{LIA}$  (see Methods section 'LIA<sub>max</sub> to 1978–87 mass balance'), and thereby the combined uncertainty estimate accounts for the scaling of the observations, and thereby incorporates the variability between observations and  $dh_{\text{solid}}$ . Thus, we regard the comparison illustrated in Extended Data Fig. 2 as a validation of our approach to derive ice-sheet-wide mass balance estimates.

**2003–2010 elevation changes from air- and space-borne laser altimetry.** To detect ice surface elevation changes from April 2003 to April 2010, which serves as the base data set from which to calculate a scale value, we use all available Ice, Cloud, and land Elevation Satellite (ICESat) GLA12 Release 31 data<sup>36</sup>. ICESat elevations have a crossover standard deviation of  $\sigma_{\text{ICESat}} = 0.2$  m (refs 37–39). Furthermore, we use all available NASA ATM flight lines<sup>35</sup> between 2003 and 2010, and NASA's Land, Vegetation, and Ice Sensor (LVIS) flight lines from 2010 (ref. 40), both of which have an uncertainty of 0.1 m. Ice surface elevation changes and associated uncertainties during the period April 2003 to April 2010 are derived in  $1 \text{ km} \times 1 \text{ km}$  cells and converted into an ice sheet surface elevation change grid ( $dh_{2003-2010}$ )<sup>18,38,41–43</sup>.

Using SMB fields from RACMO2.1/GR output<sup>11</sup> the elevation change due to firm compaction is calculated<sup>18,42</sup> and subtracted from the total elevation change ( $dh_{2003-2010}$ ), thereby yielding an elevation change due to solid ice changes ( $dh_{\text{solid}}$ ) on a  $1 \text{ km} \times 1 \text{ km}$  grid.

As part of our calculation, we divide the ice sheet into drainage basins (Extended Data Fig. 3). Here, we use the drainage basins from ref. 44 divided into sub-basins and we include additional areas around the ice sheet margin, yielding a total of 82 basins. Some areas on the southeast coast were omitted due to the lack of LIA input data, mainly caused by extensive snow cover at the time of acquisition of the aerial stereophotographs. Additionally, we use the Randolph Glacier Inventory, version 3.2 (ref. 45) to exclude glaciers not connected to the ice sheet and those only weakly connected, RGIFlag CL0 and CL1, respectively.

**Measuring LIA elevations from aerial photographs.** To detect ice surface elevation changes from  $LIA_{\max}$  to 1978–87 we use aero-triangulated vertical stereo photogrammetric imagery recorded during 1978–1987. The images were recorded between late July and mid-August from an altitude of 13,500 m to a scale of 1:150,000. They are part of a larger collection of images covering the entire ice-free part of Greenland, processed at the Anthropocene and Quaternary Research Group of the Centre for GeoGenetics, Natural History Museum of Denmark.

The aerial photographs were processed in the SOCET SET 5.6. software package written by BAE Systems using GR96 aero-analytical triangulated control points surveyed with GPS and provided by The Danish Geodata Agency<sup>46</sup>, a part of the Danish Ministry of Energy, Utilities and Climate. The processed aerial photographs allow us to survey trimlines, ice margins and moraines outlining

the  $LIA_{\max}$  in three dimensions with high accuracy. In the survey two types of points have been defined and measured: type 1, trimline or lateral moraines; and type 2, an active front.

Each of these types contains two surveyed data points:  $LIA_{\max}$  and the 1978–87 position and elevation of the ice margin (Extended Data Fig. 4). For type 1 the  $LIA_{\max}$  extent is determined from the trimline between the non-eroded and the freshly ice scoured bedrock or lateral moraines<sup>9</sup>. Type 2 determines the position of end moraines or other geomorphic evidence of recent glacially overridden landscape.

The data type distribution is illustrated in Extended Data Fig. 5a while Extended Data Fig. 5b illustrates the elevation difference at 3,003 points between  $LIA_{\max}$  and 1978–87 derived from 6,006 manual point measurements.

The elevation differences derived from the three-dimensional stereo-photogrammetric single-point survey ( $dh_{LIA}$  values) are assigned an uncertainty of 1 m as almost all systematic error affecting the triangulation of the images is eliminated<sup>9</sup>. Moreover, since the  $LIA_{\max}$  extent is mapped on the 1978–87 images we can ignore post-depositional effects on the moraines and glacial isostatic adjustment correction<sup>9</sup>.

**$LIA_{\max}$  to 1978–87 mass balance.** The ice mass balance since the  $LIA_{\max}$  is calculated by scaling  $dh_{LIA}$  values to the elevation changes between 2003 and 2010 ( $dh_{solid}$ ). We use the  $dh_{LIA}$  points at outlet glaciers of variable sizes (land- and marine-terminating) as well as other areas of the ice margin to determine the scale value ( $S_{LIA}$ ), derived as the ratio between the point-based  $dh_{LIA}$  and  $dh_{solid}$  of the closest grid cell. This implies that the shape of the ice profiles for different timestamps is not (directly) used; rather, we use  $dh$  point values that show the point-based thinning pattern along the periphery to derive the ice-sheet-wide thinning pattern. Subsequently, the  $S_{LIA}$  values found for each glacier are interpolated using the weighted mean to a regular  $1\text{ km} \times 1\text{ km}$  grid for each of the 82 calculation basins (Extended Data Fig. 3). For each grid point we predict an  $S$  value and assign an uncertainty,  $\sigma_{S(LIA\_rms)}$ , based on the root mean square of the predicted values within the basin. However, the total uncertainty  $\sigma_{S(LIA)}$  of  $S$  values has to account for the  $\sigma_{S(meth)}$  (see Methods section ‘Geometric approach to derive surface elevation changes’). Thus for each grid point  $i$  we obtain:

$$\sigma_{S(LIA)}^i = \sqrt{(\sigma_{S(LIA\_rms)}^i)^2 + (\sigma_{S(meth)}^i)^2} \quad (3)$$

Next, the elevation change between  $LIA_{\max}$  and 1978–87 are calculated by multiplying the  $S_{LIA}$  grid and the elevation change due to solid ice changes ( $dh_{solid}$ ) between 2003 and 2010:

$$dh_{LIA}^i = S_{LIA}^i dh_{solid}^i \quad (4)$$

where  $i$  represents each cell on a regular  $1\text{ km} \times 1\text{ km}$  grid. By using  $dh_{solid}$ , which includes changes in elevation due to firn compaction, we thereby obtain estimates of the mass balance.

To each value of  $dh_{LIA}$  we assign uncertainty as follows:

$$\sigma_{dh_{LIA}}^i = dh_{LIA}^i \sqrt{\left(\frac{\sigma_{S(LIA)}^i}{S_{LIA}^i}\right)^2 + \left(\frac{\sigma_{dh_{solid}}^i}{dh_{solid}^i}\right)^2} \quad (5)$$

The calculation allows us to ignore the actual timing of the maximum extent during  $LIA$  because it is mapped on the 1978–87 images, thereby making it directly applicable to derive ice net mass balance between  $LIA_{\max}$  and 1978–87. However, to obtain a rate we assign 1900 as a Greenland-wide time stamp of when the glaciers started to retreat following the  $LIA$ , although we note that there is regional and local variability<sup>9,12,13</sup>, and use 1983 as the average year of the aerial imagery. **Photogrammetric DEM 1978–87.** We produced a  $25\text{ m} \times 25\text{ m}$  digital elevation model (DEM1978/87) using the vertical stereo photogrammetric imagery recorded during 1978–1987 following a standard approach<sup>18,22</sup>. The DEM is processed into WGS84 ellipsoid heights, directly comparable to ICESat, ATM, and LVIS data.

Our validation methodology is based upon co-registration methods that relate the three-dimensional co-registration vector between two elevation surfaces to terrain slope  $\alpha$  and aspect  $\psi$  (refs 47 and 48). The co-registration parameters are determined by robust least-squares minimizations of stable terrain elevation changes between the DEM tiles and ICESat<sup>36</sup> ( $dh$ ) using:

$$dh = a \cos(b - \psi) \tan(\alpha) + c \quad (6)$$

where  $a$  and  $b$  is the magnitude and direction, respectively, of the horizontal co-registration vector and  $c$  is the mean vertical bias between the two elevation data sources.

We perform the co-registration on a  $50\text{ km} \times 50\text{ km}$  grid over all the DEM. All slopes less than  $5^\circ$  are removed and a curvature filter is applied to remove regions where resolution variation between the data sets may cause spurious elevation differences.

The co-registration parameters are generally less than 15 m horizontally and less than 10 m vertically. At the  $1\sigma$  confidence level, the aero-photogrammetric DEM has an accuracy of 10 m horizontally and 6 m vertically while the precision is better than 4 m (Extended Data Fig. 6). We note that the 6 m vertical accuracy of the DEM is different from the 1 m uncertainty related to  $dh_{LIA}$  values obtained in the three-dimensional stereo-photogrammetric single-point survey (see Methods section ‘ $LIA_{\max}$  to 1978–87 mass balance’).

**1978–87 to 2003 mass balance.** The mass balance between 1978–87 and 2003 is determined using the same approach as outlined for calculating the  $LIA_{\max}$  to 1978–87 mass balance but with different input data. We use ATM data from 2010 (ref. 35), supplemented with 2009 ICESat data<sup>36</sup> to fill in gaps, to determine the mass balance between 1978–87 and 2010. Subsequently, we subtract the derived mass balance between 2003 and 2010 to determine the 1978–87 to 2003 mass balance.

The merged ATM and ICESat data cover outlet glaciers of variable size and termination regime. At these data points, elevations from the 1978–87 DEMs are extracted, although we remove interpolated DEM surfaces using a reliability mask<sup>18</sup>, an output produced during DEM production. The point-based difference between the ATM/ICESat measured surface elevation and the DEM elevation is  $dh_{80s-10}$ . To accommodate issues related to large differences between simulated and predicted surface profiles (see Methods section ‘Geometric approach to derive surface elevation changes’) we use point observations only up-glacier from the terminus, though the distance varies for individual outlet glaciers with the location of available ATM and ICESat data.

Next, we derive the  $S_{80s-10}$  value as the ratio between the point-based  $dh_{80s-10}$  and a  $dh_{solid}$  value extracted from the  $1\text{ km} \times 1\text{ km}$  grid using bilinear interpolation between grid cells. The  $S_{80s-10}$  values are subsequently interpolated using a weighted mean to a regular  $1\text{ km} \times 1\text{ km}$  grid. Thus for each grid point we predict a  $S$  value and assign an uncertainty,  $\sigma_{S(80s-10\_rms)}$ , based on the root mean square of the predicted values within the basin. However, the total uncertainty  $\sigma_{S_{80s-10}}$  of  $S$  values has to account for the  $\sigma_{S(meth)}$  (see Methods section ‘Geometric approach to derive surface elevation changes’). Thus, for each grid point  $i$  we obtain:

$$\sigma_{S(80s-10)}^i = \sqrt{(\sigma_{S(80s-10\_rms)}^i)^2 + (\sigma_{S(meth)}^i)^2} \quad (7)$$

The elevation change between 1978–87 and 2010 is calculated by multiplying the  $S_{80s-10}$  grid and the  $dh_{solid}$  (2003–2010) grid:

$$dh_{80s-10}^i = S_{80s-10}^i dh_{solid}^i \quad (8)$$

where  $i$  represents each cell on a regular  $1\text{ km} \times 1\text{ km}$  grid.

To each value of  $dh_{80s-10}$  we assign an uncertainty of:

$$\sigma_{dh_{80s-10}}^i = dh_{80s-10}^i \sqrt{\left(\frac{\sigma_{S(80s-10)}^i}{S_{80s-10}^i}\right)^2 + \left(\frac{\sigma_{dh_{solid}}^i}{dh_{solid}^i}\right)^2} \quad (9)$$

Subsequently, we subtract  $dh_{solid}^i$  from  $dh_{80s-10}^i$  to determine the mass balance from 1978–87 to 2003.

**Uncertainties and conservative mass balance estimates.** For the entire ice sheet (or individual basins) we calculate the uncertainty of the mass balance estimates during  $LIA_{\max}$  to 1978–87 and 1978–87 to 2003 as:

$$\sigma_{POI} = \sum_{i=1}^n \sigma_{dh_{POI}}^i \quad (10)$$

where  $\sigma_{dh_{POI}}^i$  is the uncertainty of each grid point during the period of interest ( $LIA_{\max}$  to 1978–87 or 1978–87 to 2003) derived from equation (5) or equation (9) and  $n$  is the number of points covering the basin, region, or entire ice sheet being considered.

We regard the derived mass balance estimates between  $LIA_{\max}$  and 1978–87 as conservative for a number of reasons. First, when calculating mass balance we are limited by the spatial extent of our ice mask, which implies that mass loss between the boundary of the ice mask (based on the ice extent derived from the 1978–87 aerial images) and the maximum extent of the glaciers during the  $LIA$  is not included. This zone of non-included ice loss is largest near marine-terminating glaciers. For example Jakobshavn Isbræ retreated by about 20 km between  $LIA_{\max}$  and 1978–87, while Kangerlussuaq Glacier and Midgaard Glacier retreated by about 12 km and about 20 km, respectively, during the same period. Here, the outer parts of the glaciers may have been afloat during the  $LIA$  and would already then



have contributed to GMSL rise, while only mass loss up-glacier from the  $LIA_{\max}$  grounding line would contribute to post-LIA sea level rise. As the extent of the ice mask is not identical to the  $LIA_{\max}$  grounding line we cannot capture the mass loss between these two.

Second, owing to the lack of LIA data on the southeast some areas are excluded, and so are glaciers not connected to the ice sheet and those only weakly connected, RGIFlag CL0 and CL1, respectively<sup>45</sup>. This may lead to a smaller mass balance estimate for the different periods; for example we estimate a mass loss of  $186.4 \pm 18.9 \text{ Gt yr}^{-1}$  for the period 2003–2010, while using the same ice mask as ref. 18 we arrive at a mass loss of  $250.1 \pm 21.2 \text{ Gt yr}^{-1}$ .

Third, propagation of thinning at the ice margin towards the interior is not incorporated in the present model, as we use a scaling approach based on point-based thinning observations at the periphery of the ice sheet to estimate the mass balance. Model experiments suggest that mass loss at lower elevations would propagate inland and cause interior thinning on decadal timescales and continue inland even if mass loss at the ice margin ceases<sup>49</sup>. For all three periods we calculate a mass gain in the interior of the ice sheet, which since 1993 has also been identified by others<sup>50–52</sup>. Verifying mass gain in the interior for the period  $LIA_{\max}$  to 1978–87 (1983) is difficult because ice-core-derived estimates of ice surface elevation changes are associated with vertical uncertainties of about 70 m (ref. 53). Excluding the interior mass gain during  $LIA_{\max}$ –1983 yields a mass loss of  $7,712 \pm 1323 \text{ Gt}$  ( $92.9 \pm 15.9 \text{ Gt yr}^{-1}$ ) relative to the conservative estimate of  $6,233 \pm 2436 \text{ Gt}$  ( $75.1 \pm 29.4 \text{ Gt yr}^{-1}$ ).

Even though the individual contribution of the abovementioned assumptions to mass loss may be considered minor, the combined effects may be considerable. However, given the limitations in the present model configuration and lack of observations to constrain the behaviour of the interior mass balance, we favour the conservative estimate presented in this paper and emphasize that it should be regarded as a minimum contribution from the GIS to GMSL rise during the period  $LIA_{\max}$  to 1978–87 (1983).

**SMB modelling.** The near-surface air temperature  $T$  and the land-ice SMB (that is, total precipitation minus total sublimation minus runoff) reconstruction of ref. 10, spanning 1840–2012, is calibrated to RACMO2.1/GR output<sup>11</sup>. The calibration is important because SMB fields from RACMO2.1/GR are used as input to convert total elevation during 2003–2010 ( $dh_{2003-2010}$ ) into elevation change due to solid-ice changes ( $dh_{\text{solid}}$ ) using a firn-compaction model (see Methods section ‘2003–2010 elevation changes from air- and space-borne laser altimetry’). Thus, because we use  $dh_{\text{solid}}$  to calculate mass balance estimates during  $LIA_{\max}$ –1983 and 1983–2003, it is critical that the two SMB models are comparable when assessing the components of the twentieth-century mass balance. Furthermore, owing to sharply decreasing ice core data availability after 1999, from which snow accumulation is derived in the SMB model of ref. 10, the model incorporates precipitation fields from RACMO2.1/GR. The calibration of  $T$  and the SMB components excluding snow accumulation employs a 53-year overlap period (1960–2012), whereas snow accumulation is calibrated during the 1960–1999 period. The calibration employs linear regression coefficients at each 5-km grid cell that match the multi-year average of the reconstruction with that from RACMO2.1/GR. Prior to calibration the RACMO2.1/GR data are resampled/reprojected from their native  $0.1^\circ$  ( $\sim 11 \text{ km}$ ) grid to the 5-km grid employed by ref. 10. A 5%–8% correction is applied in SMB totals to account for the 5-km polar stereographic grid cell area variation with latitude.

Refinements are applied to the original SMB reconstruction<sup>10</sup> as follows. (1) Values are now estimated over all land, sea, and ice within the domain, rather than over only ice. (2) A physically based meltwater retention scheme<sup>54</sup> replaces the original simpler approach. (3) Multiple stations now contribute to the  $T$  value for each given month and grid cell within the domain, rather than employing the single highest-correlating station. (4) The RACMO2.1/GR data used for calibration have a higher native resolution ( $\sim 11 \text{ km}$ ) than the Polar MM5 data ( $\sim 24 \text{ km}$ ) used to calculate the original SMB reconstruction. (5) The SMB reconstruction now extends to 2012, rather than 2010. (6) The ice-core-derived annual accumulation rates are divided into monthly temporal resolution by weighting the monthly fraction of annual accumulation after the 1960–2012 average RACMO2.1/GR seasonal distribution at each grid cell.

Absolute uncertainty for the revised SMB estimates from ref. 10 is estimated by comparing against field data. *In situ* annual ablation rates ( $n = 208$ ), spanning 1985–1992, yield an ablation root-mean-square error of 35%. This is analogous to an *in situ* comparison with RACMO2.1/GR. Comparison between revised SMB estimates from ref. 10 (or RACMO2.1/GR) with ice-core-derived net accumulation time series from 86 sites<sup>55</sup> yields a 30% accumulation root-mean-square error.

A fundamental assumption is that the calibration regression factors (slope and intercept), derived on a grid cell basis during 1960–2012 versus ice cores, meteorological station temperatures, and with RACMO2.1/GR, are stationary in time. Testing this, we find that over the 53-year overlap period (1960–2012) cumulative

SMB anomalies drift between the reconstruction and RACMO2.1/GR by up to 600 Gt as compared to a total mass flux of 24,000 Gt, suggesting a drift uncertainty of 2.5%. In the pre-1960 period, cumulative uncertainty may be larger.

**Temporal variability of the mass balance.** To assess the variability of the mass balance during the twentieth century we use an approach similar to that of other studies<sup>3,5,24</sup>. Here iceberg discharge is estimated using a linear regression between reconstructed meltwater runoff from revised SMB estimates from ref. 10 and estimates of ice-sheet-wide iceberg discharge, spanning 2000–2012 (ref. 21). We find a peak correlation ( $r = 0.87$ ;  $P < 0.01$ ; degrees of freedom = 12) between annual iceberg discharge  $D$  and six-year mean meltwater runoff  $R_6$ , calculated from the five years preceding, and including a given year.  $D_M$  is modelled ice discharge in gigatonnes per year and is calculated as follows:

$$D_M = 0.766R_6 + 266 \quad (11)$$

This is similar to employing a correlation between five-year lagging meltwater runoff and annual iceberg discharge<sup>24</sup>. We use the discharge estimates of ref. 21, though we note that the estimates generally lie within uncertainties of other studies<sup>3,5,6</sup> except those of ref. 24 (Extended Data Fig. 7). Uncertainties related to the temporal mass balance method are calculated using Monte Carlo simulation (see Methods section ‘Estimating uncertainties using Monte Carlo simulation’).

**Estimating uncertainties using Monte Carlo simulation.** We implement a Monte Carlo uncertainty approach that accounts for the interaction of uncertainties in mass balance components<sup>5</sup>. The residual root-mean-square differences between revised SMB estimates from ref. 10 and RACMO2.1/GR are increased by 50% to form a conservative uncertainty estimates given that the absolute uncertainty may be larger than the calibration root-mean-square difference. The post-calibration root-mean-square difference for runoff is increased by 50% yielding an assumed conservative uncertainty of 24.9%. That for accumulation is 8.0% and that for SMB is 22.5%. These are relative uncertainties between RACMO2.1/GR and revised SMB estimates from ref. 10. Absolute uncertainty is evaluated relative to field data.

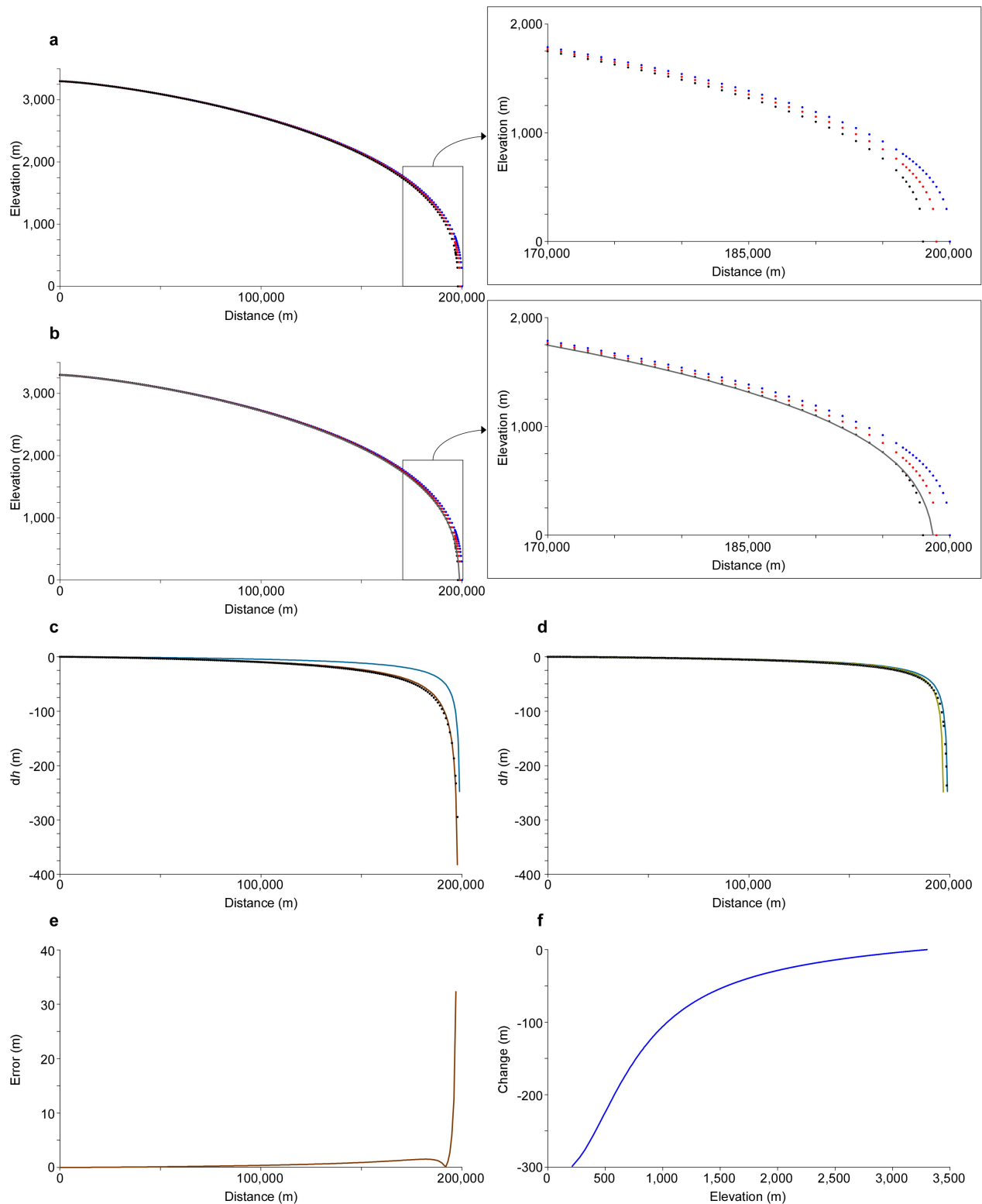
Because iceberg discharge is a function of runoff, the runoff uncertainty is propagated through Eq. (11) to estimate iceberg discharge uncertainty. The temporal mass balance uncertainty is estimated as  $78 \text{ Gt yr}^{-1}$ . Extended Data Fig. 8 shows the Monte Carlo simulation for the temporal mass balance expressed as cumulative eustatic sea level change during 1840–2012.

**Data.** We use aero-triangulated vertical stereo photogrammetric imagery recorded during 1978–1987 to manually map the former ice extent during the  $LIA_{\max}$ . Raw imagery was made available for research purposes by The Danish Geodata Agency, a part of the Danish Ministry of Energy, Utilities and Climate. The derived products used in this study such as orthophotos and the DEM are available in GeoTiff format upon request to the corresponding author. Moreover, we use all available ICESat GLA12 Release 31 data<sup>36</sup> (<https://nsidc.org/data/icesat/data.html>) and all available NASA ATM flight lines<sup>35</sup> between 2003 and 2010 (<http://nsidc.org/data/blatm2> and <https://nsidc.org/data/ilatm2>) and NASA’s LVIS flight lines<sup>40</sup> from 2010 (<https://nsidc.org/data/ilvis2>). Information on SMB data from RACMO2.1/GR<sup>11</sup> is available at <http://www.projects.science.uu.nl/iceclimate/models/greenland.php>, while information on SMB is available from ref. 10. To model ice discharge we use ice discharge estimates from ref. 21.

**Code availability.** Data analyses have been performed using the SOCET SET 5.6 software package (written by BAE Systems), ArcGIS10.1 (written by Esri Inc.), and custom-built routines for Python, Matlab and Fortran. The codes are not available.

31. Glasser, N. F., Harrison, S., Jansson, K. N., Anderson, K. & Cowley, A. Global sea-level contribution from the Patagonian Icefields since the Little Ice Age maximum. *Nature Geosci.* **4**, 303–307 (2011).
32. Carrivick, J. L., Davies, B. J., Glasser, N. F., Nývlt, D. & Hambrey, M. J. Late-Holocene changes in character and behaviour of land-terminating glaciers on James Ross Island, Antarctica. *J. Glaciol.* **58**, 1176–1190 (2012).
33. Cuffey, K. M. & Paterson, W. S. B. *The Physics of Glaciers* (Elsevier, 2010).
34. Rignot, E. & Mouginot, J. Ice flow in Greenland for the International Polar Year 2008–2009. *Geophys. Res. Lett.* **39**, L11501 (2012).
35. Krabill, W. B. *IceBridge ATM L2 Lcassn Elevation, Slope, and Roughness, [2003–2010] data set* <http://nsidc.org/data/blatm2> (National Snow and Ice Data Center, 2014).
36. Zwally, H. J. et al. *GLAS/ICESat L2 Antarctic and Greenland Ice Sheet Altimetry Data V031 data set* <https://nsidc.org/data/icesat/data.html> (National Snow and Ice Data Center, 2011).
37. National Snow and Ice Data Center. *ICESat: Description of Data Releases* data set [http://nsidc.org/data/icesat/data\\_releases.html](http://nsidc.org/data/icesat/data_releases.html) (2011).
38. Howat, I. M., Smith, B. E., Joughin, I. R. & Scambos, T. A. Rates of southeast Greenland ice volume loss from combined ICESat and ASTER observations. *Geophys. Res. Lett.* **35**, L17505 (2008).
39. Pritchard, H. D., Arthern, R. J., Vaughan, D. G. & Edwards, L. A. Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. *Nature* **461**, 971–975 (2009).

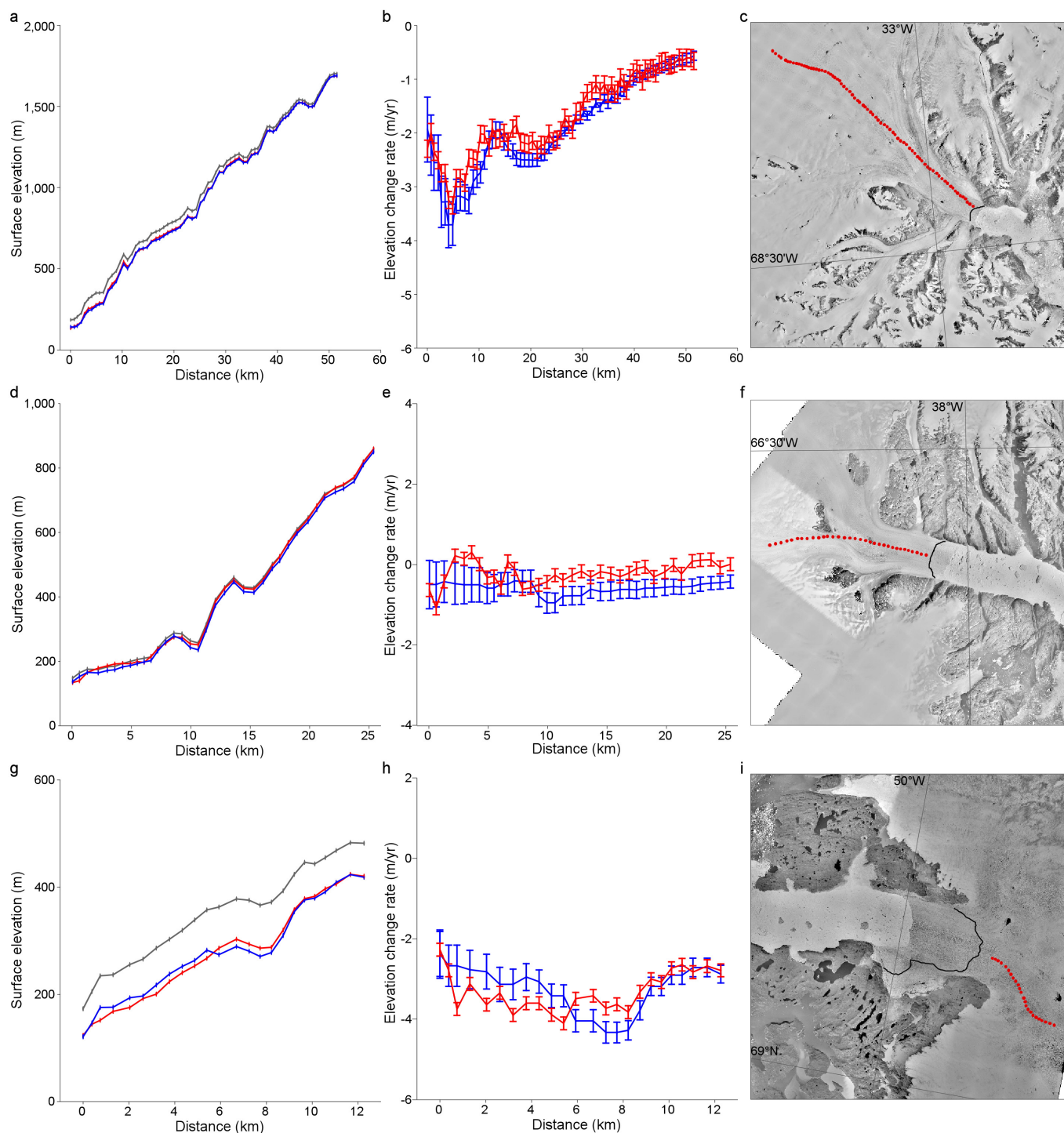
40. Blair, B. & Hofton, M. *IceBridge LVIS L2 Geolocated Ground Elevation and Return Energy Quartiles, [2010] data set*. <http://nsidc.org/data/ilvis2.html> (National Snow and Ice Data Center, 2010).
41. Ewert, H., Groh, A. & Dietrich, R. Volume and mass changes of the Greenland ice sheet inferred from ICESat and GRACE. *J. Geodyn.* **59–60**, 111–123 (2012).
42. Kjeldsen, K. K. *et al.* Improved ice loss estimate of the northwestern Greenland ice sheet. *J. Geophys. Res. Solid Earth* **118**, 698–708 (2013).
43. Smith, B. E., Fricker, H. A., Joughin, I. R. & Tulaczyk, S. An inventory of active subglacial lakes in Antarctica detected by ICESat (2003–2008). *J. Glaciol.* **55**, 573–595 (2009).
44. Rignot, E. & Kanagaratnam, P. Changes in the velocity structure of the Greenland Ice Sheet. *Science* **311**, 986–990 (2006).
45. Arendt, A. *et al.* *Randolph Glacier Inventory—A Dataset of Global Glacier Outlines Version 3.2*, <http://www.glims.org/RGI/> (Global Land Ice Measurements from Space, 2013).
46. Danish Geodata Agency. *Ground control for 1:150,000 scale aerials, Greenland* <http://gst.dk/emner/landkort-topografi/groenland/ground-control-greenland/> (GST, 2013).
47. Nuth, C. & Kääb, A. Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change. *Cryosphere* **5**, 271–290 (2011).
48. Kääb, A. *Remote Sensing of Mountain Glaciers and Permafrost Creep. Schriftenreihe Physische Geographie* (Univ. Zürich, Department of Geography, 2005).
49. Wang, W., Li, J. & Zwally, H. J. Dynamic inland propagation of thinning due to ice loss at the margins of the Greenland ice sheet. *J. Glaciol.* **58**, 734–740 (2012).
50. Krabill, W. B. *et al.* Greenland Ice Sheet: increased coastal thinning. *Geophys. Res. Lett.* **31**, L24402 (2004).
51. Sasgen, I. *et al.* Timing and origin of recent regional ice-mass loss in Greenland. *Earth Planet. Sci. Lett.* **333–334**, 293–303 (2012).
52. Hurkmans, R. T. W. L. *et al.* Time-evolving mass loss of the Greenland Ice Sheet from satellite altimetry. *Cryosphere* **8**, 1725–1740 (2014).
53. Lecavalier, B. S. *et al.* Revised estimates of Greenland ice sheet thinning histories based on ice-core records Greenland. *Quat. Sci. Rev.* **63**, 73–82 (2013).
54. Pfeffer, W. T., Meier, M. F. & Illangasekare, T. H. Retention of Greenland runoff by refreezing: implications for projected future sea level change. *J. Geophys. Res.* **96**, 22117–22124 (1991).
55. Box, J. E. *et al.* Greenland ice sheet mass balance reconstruction. Part I: Net snow accumulation (1600–2009). *J. Clim.* **26**, 3919–3934 (2013).
56. Andersen, M. L. *et al.* Basin-scale partitioning of Greenland ice sheet mass balance components (2007–2011). *Earth Planet. Sci. Lett.* **409**, 89–95 (2015).
57. Weidick, A. Historical fluctuations of calving glaciers in South and West Greenland. *Rapp. Grønlands Geol. Unders.* **161**, 73–79 (1994).
58. Weidick, B. A. Neoglacial glaciations around Hans Tausen Iskappe, Peary Land, North Greenland. *Medd. Grøn. Geosci.* **39**, 5–26 (2001).



**Extended Data Figure 1 | The  $dh$  calculation scheme.** **a**, Three simulated ice surface profiles based on Glen's flow law, each representing time steps t1 (blue dots), t2 (red dots), and t3 (black dots). **b**, The same profiles as in **a** supplemented with the predicted profile  $h_{pre-t3}$  (grey line) derived using an  $S$  value of 2.2. The figure shows agreement between the profile  $h_{t3}$  and  $h_{pre-t3}$ ; hence, if we know the elevation change during one period (for example, t1 and t2), then it is possible to obtain the elevation change during another period (for example, t1 and t3) by multiplying with a constant  $S$ . **c**, The elevation changes between t1 and t2 ( $dh_{t1t2}$ , blue line) and between t1 and t3 ( $dh_{t1t3}$ , brown line). The black dots are the elevation changes between t1 and the predicted surface profile  $h_{pre-t3}$  derived using the elevation change between t1 and t2 and an  $S$  value of 2.2. The predicted difference

( $dh_{t1t3\_pre}$ ) between t1 and t3 is derived from  $dh_{t1t2}$  and a constant, implying that the surface profile at t1 is part of both the input and of the output. **d**,  $dh_{t1t2}$  (blue line), the elevation changes between t3 and t4 ( $dh_{t3t4}$ , dark green line), and the predicted  $dh_{t3t4}$  ( $dh_{t3t4\_pre}$ , black dots), which is derived using  $dh_{t1t2}$  and an  $S$  value of 1.2; thus none of the ice surface profiles are part of both the input and output. If both  $dh_{t1t2}$  and  $dh_{t3t4}$  are known then  $S$  can be derived as the ratio between the observations. **e**, The uncertainty between the profile  $h_{t3}$  and  $h_{pre-t3}$  using a constant  $S$ . Generally the differences are small, though they increase near the margin. **f**, The elevation change between two time steps as a function of elevation. Changes are largest at lower elevation and become close to 0 at  $h > 2,500$  m.

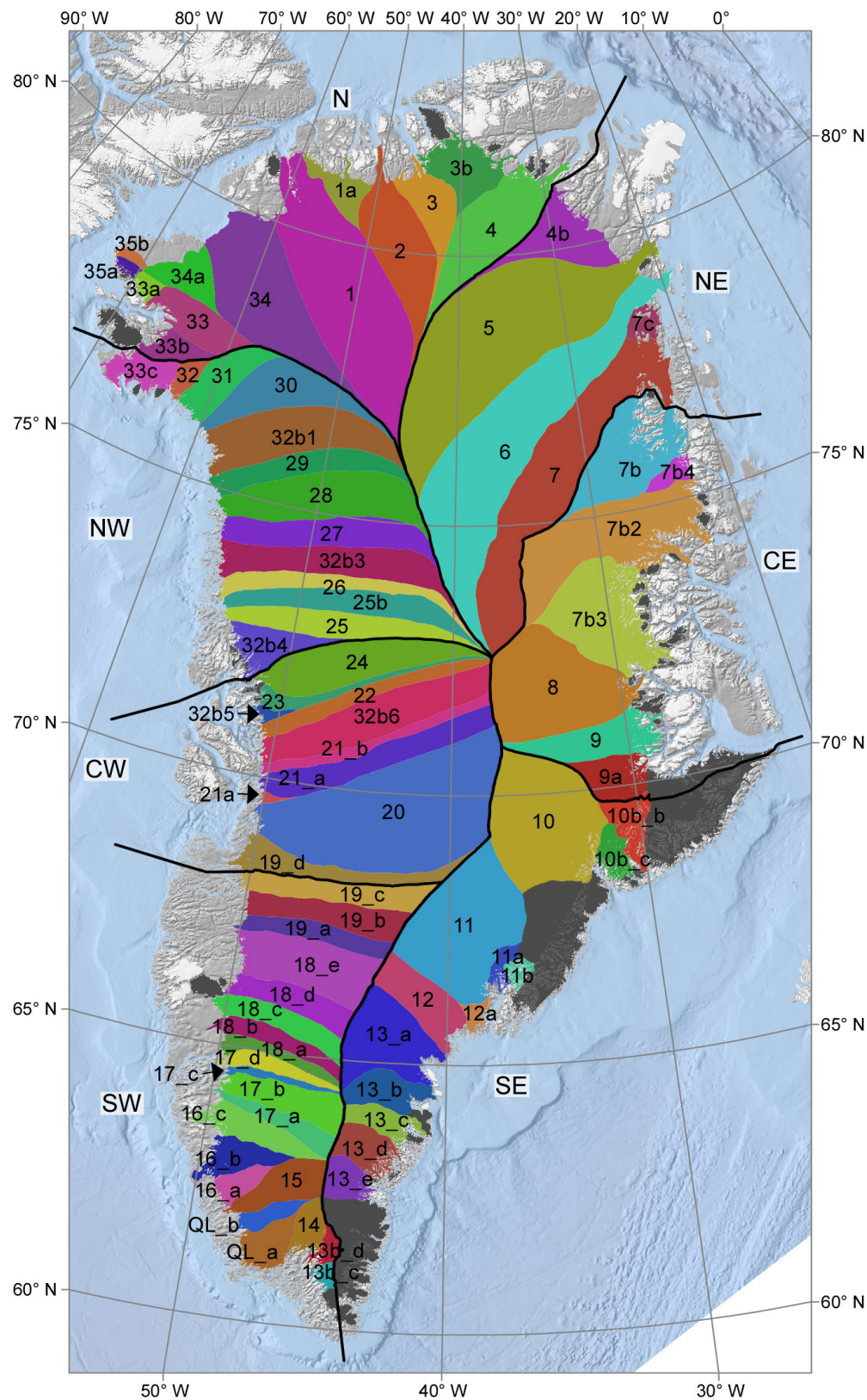




### Extended Data Figure 2 | Validation of the scaling approach.

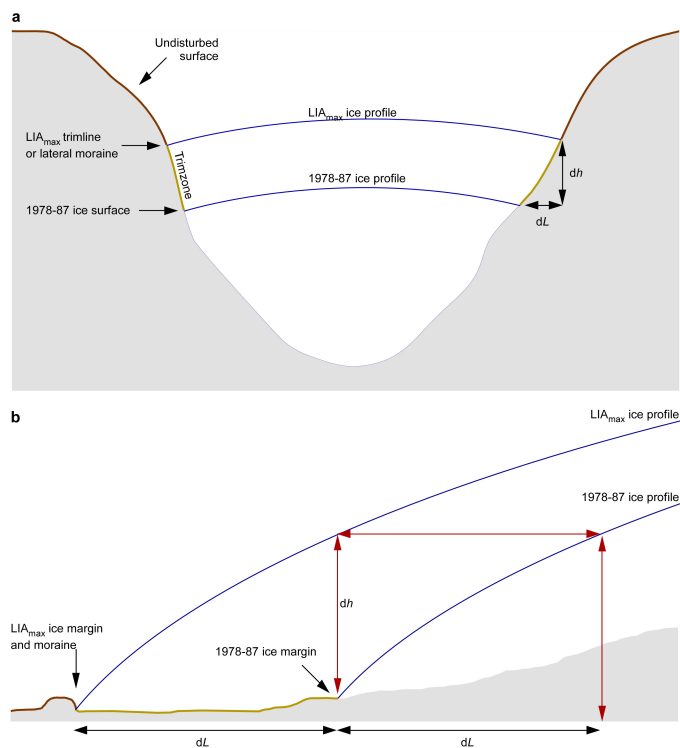
**a**, Elevation profiles of Kangerlussuaq Glacier in southeast Greenland from the 1981 DEM (grey line), 2003 ATM data (red line), and the predicted surface profile (blue line) in 2003, derived using the scaling approach based on local scale values and the 2003–2010 elevation changes ( $dh_{\text{solid}}$ ). (For a more complete description of the approach using observations see Methods section ‘LIA<sub>max</sub> to 1978–87 mass balance’). **b**, The elevation change rate between the observed 2003 surface profile (red) and the predicted 2003 surface profile (blue) relative to the 1981 DEM. The blue vertical lines denote uncertainty estimates that include an uncertainty related to the scaling approach, an error related to observed changes during 2003–2010, and an uncertainty related to the scaling of point-based observations. The red vertical lines denote an uncertainty associated with

the observed elevation changes during 1981–2003 and includes combined errors of the measured height derived from stereo photogrammetric DEM and 2003 ATM data. **c**, A 1981 orthophoto of Kangerlussuaq Glacier with 2003 ATM data (red dots) and the May 2003 glacier front (black line). **d–f** and **g–i** illustrate the same as **a–c** for Helheim Glacier and Jakobshavn Isbræ, respectively. However, for Jakobshavn Isbræ the DEM and orthophoto is from 1985. Note the different scales for each of the glaciers. Comparing the elevation change rates derived from the scaling approach and those directly from the observations, we find good agreement as the error bars overlap. Thus, we regard the illustrated comparison as a validation of our method of deriving ice-sheet-wide mass balance estimates.



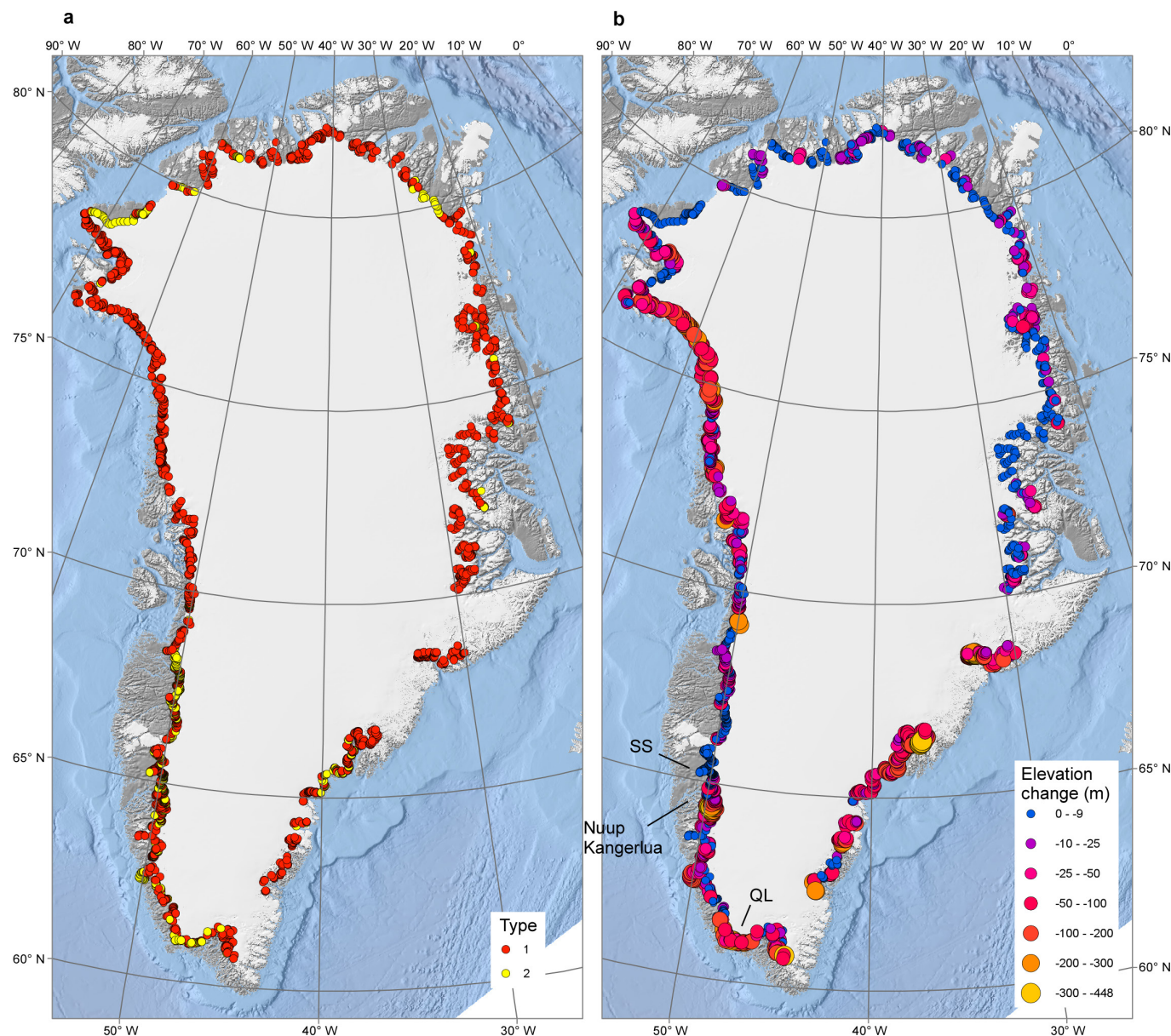
**Extended Data Figure 3 | GIS calculation basin subdivision.** Calculation basins modified from ref. 44 to include slower-moving areas of the ice sheet. Note that three areas on the southeast coast have been omitted due to an insufficient number of LIA to 1978–87 data points caused by

extensive snow cover on the vertical images. The total ice mask covers 1,647,907 km<sup>2</sup>. The additional areas included in the ice mask used by ref. 18 are shown in dark grey and in total the ice mask covers 1,739,564 km<sup>2</sup>.



**Extended Data Figure 4 | Mapping elevation changes during LIA<sub>max</sub> to 1978–87.** **a**, Type 1 points are placed at the trimline or lateral moraine marking the LIA<sub>max</sub> position and at the 1978–87 ice surface perpendicular to the flow direction, and as we assume that the cross-section profile of the glacier is the same during the LIA<sub>max</sub> and 1978–87 then the vertical difference  $dh$  is the thinning at this location. This approach is the same as used by ref. 9. **b**, For Type 2 points we assume that the longitudinal shape of the glacier is the same during the LIA<sub>max</sub> as in 1978–87. Points are placed at the LIA<sub>max</sub> margin and at the 1978–87 margin, and assuming a longitudinal profile that does not change over time, the distance  $dL$  is used to find the vertical difference between the 1978–87 point and a point on the glacier at a distance of  $dL$  following the same flowline. Points for glaciers receding on steep slopes have been discarded.

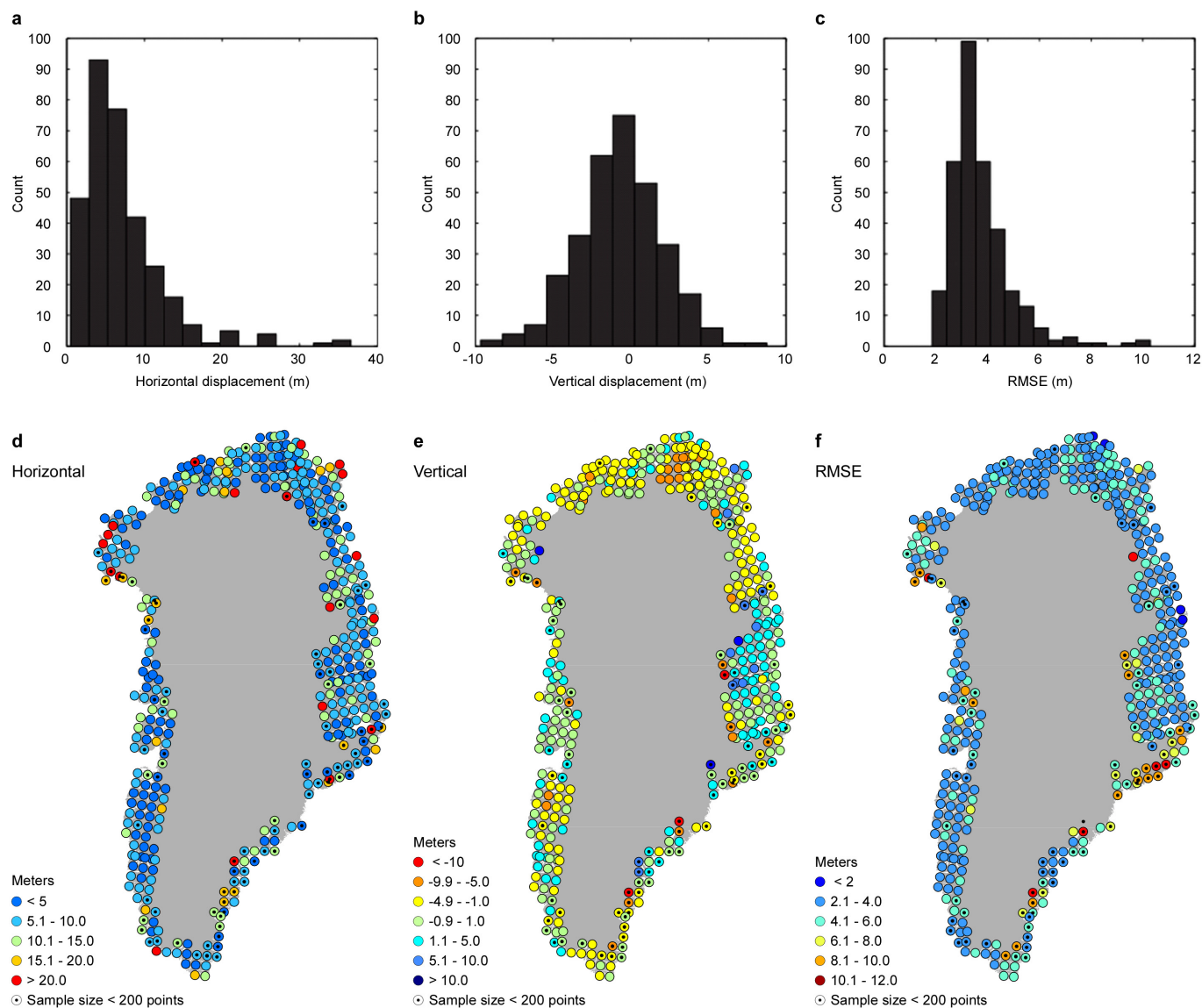




### Extended Data Figure 5 | Distribution and values of $dh_{LIA}$ points.

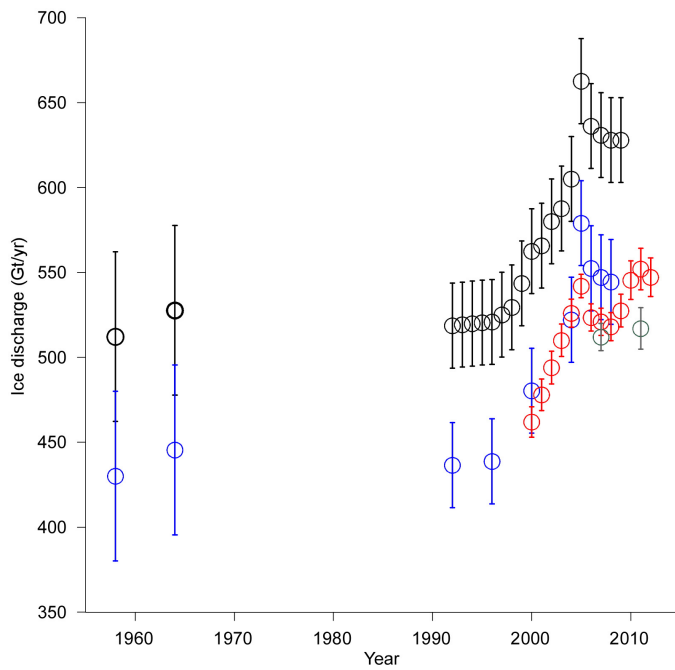
**a**, Distribution of the two point types used to determine thinning between  $LIA_{max}$  and 1978–87. **b**, From the type 1 and type 2 points, net elevation change  $dh_{LIA}$  is measured based on 3,003 point measurements from the  $LIA_{max}$  to 1978–87. Of the 3,003 pairs—that is, 6,006 point measurements—2,476 are measured as type 1 and 527 are type 2. The majority of the type 2 points are found along the land-terminating and slower-moving parts of the ice sheet, whereas type 1 points are found in valleys through which the ice flows and on nunataks.  $dh_{LIA}$  values range between zero and  $-448$  m (a negative value implies thinning). The largest  $dh_{LIA}$  values are found along the major marine-outlet glaciers along the northwest and southeast coast and along the rim of the Qassimiut lobe (QL), while in contrast the lower  $dh_{LIA}$  values are found along the slower-moving margins of the typically land-terminating ice sheet. In some areas around the ice sheet no trimlines are visible and/or the ice margin is in contact with the  $LIA$  moraines. Analysis of glacier front positions for outlet glaciers in the north, central west, southwest, and south using historical aerial photographs from the 1930s and onwards<sup>15,57,58</sup> suggest

that a few outlet glaciers, primarily land-terminating, have been stable or advanced since the  $LIA$ . In the northwest, central west, and southwest snow cover on the 1978–87 vertical aerial images is generally limited, which eases the distinction between freshly eroded bedrock, newly deposited glacial sediment, and non-eroded vegetated terrain surfaces. This supports the notion that if no trimline is visible on the photographs, then the ice margin is at an advanced and stable stage. Hence, the  $dh_{LIA}$  and  $dL_{LIA}$  values for points are zero. An example of a glacier that has advanced during the twentieth century is the Saqqap Sermia (SS)<sup>57</sup> in the Nuup Kangerlua (Godthåbsfjord) complex in southwest Greenland. Here no trimlines are visible along the valley and the boundary between ice and vegetation cover is only interrupted by small meltwater channels, and at the glacier front no end moraines are visible on the meltwater plain. In the present setup we are not able to assign any post- $LIA$  mass gain; however, as only a limited number of outlet glaciers have advanced and exceeded the  $LIA$  front position during the twentieth century we regard this mass gain as negligible relative to the ice-sheet-wide mass loss.



**Extended Data Figure 6 | Horizontal and vertical displacements in aero-photogrammetric DEM.** **a–c**, Histograms of the horizontal (**a**) and vertical (**b**) co-registration displacements for each 50 km × 50 km grid cell show that the aero-photogrammetric DEM compilation is generally accurate to within 10 m horizontally and 6 m vertically with a precision greater than 4 m ( $1\sigma$  confidence level) (**c**). **d–f**, The horizontal

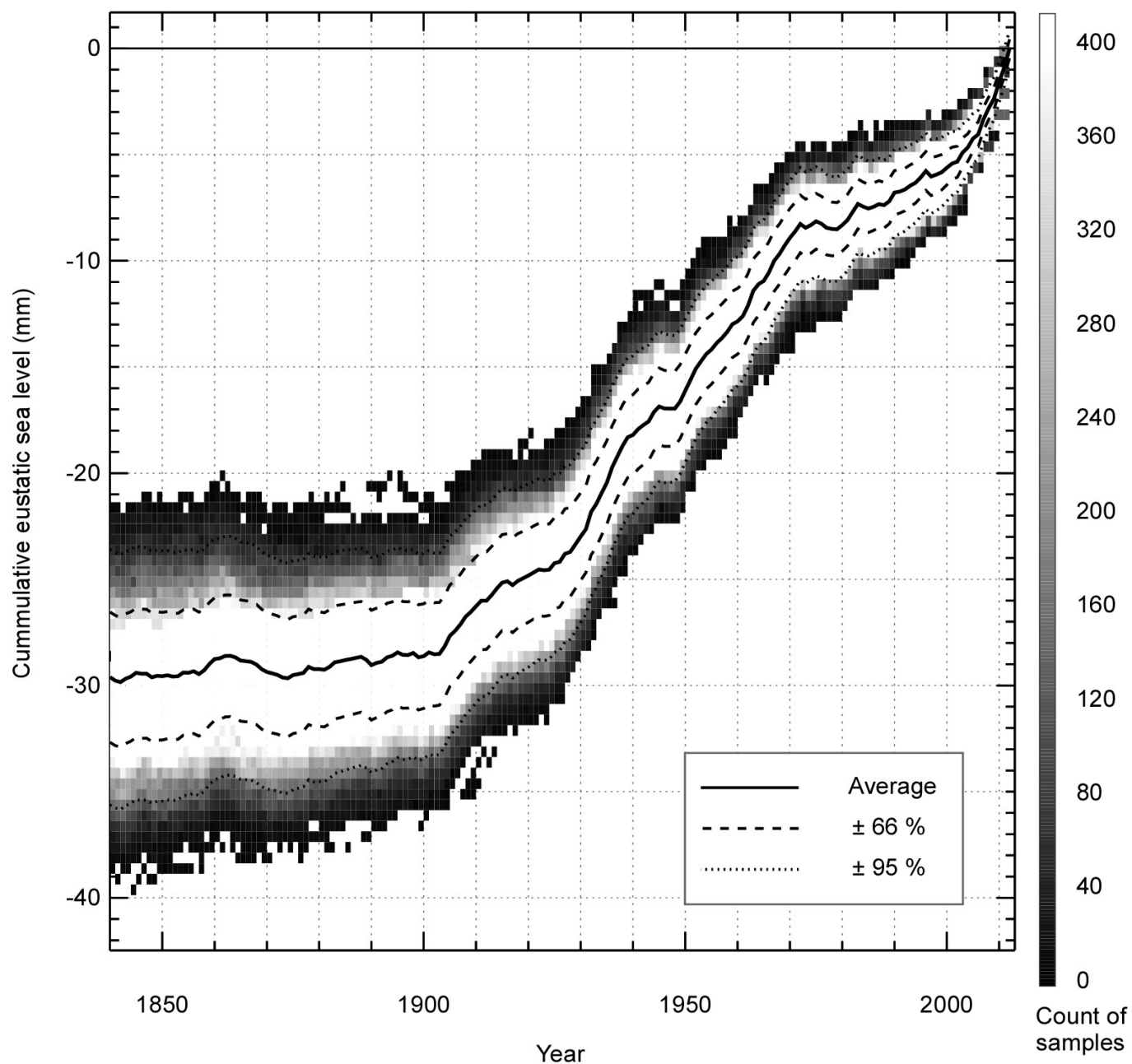
(**d**) and vertical (**e**) components of the co-registration vectors between 50 km × 50 km sections of the aero-photogrammetric DEM compilation and ICESat laser altimetry are plotted with the root-mean-square error of stable terrain differences after adjusting for the three-dimensional mis-registration (**f**).



**Extended Data Figure 7 | Estimates of ice-sheet-wide iceberg discharge.**

Ice discharge estimates and associated errors (vertical bars) from ref. 24 (black), ref. 3 (blue), ref. 21 (red), and ref. 56 (grey). We note that the used discharge estimates of ref. 21 are  $15 \text{ Gt yr}^{-1}$  greater than those of ref. 56,  $30 \text{ Gt yr}^{-1}$  less than those of ref. 3, and  $110 \text{ Gt yr}^{-1}$  less than those of ref. 24. Such discrepancies are attributed to differences in data availability and assumptions used for filling gaps or the method used to correct for SMB between the inland flux gates and the grounding lines<sup>21</sup>.





**Extended Data Figure 8 | Temporal variability of the mass balance expressed as cumulative eustatic sea level rise.** Reconstructed temporal mass balance during the period 1840–2012 derived using revised SMB

estimates from ref. 10 and modelled ice discharge, calculated as a function of six-year average runoff. The uncertainty is assessed from a Monte Carlo simulation using 4,000 samples for each year.

# Grassland biodiversity bounces back from long-term nitrogen addition

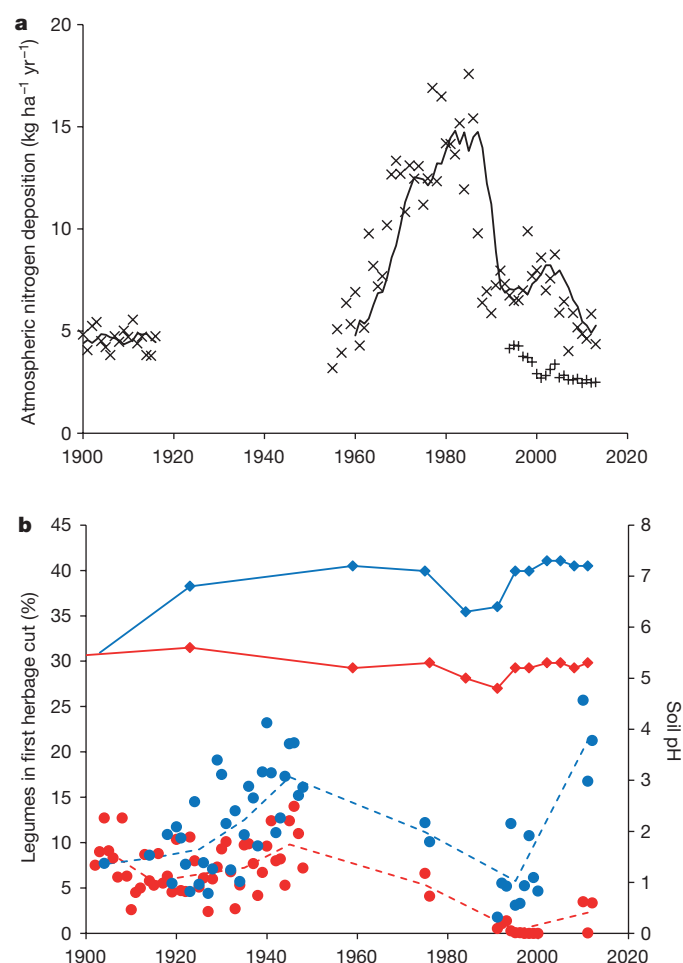
J. Storkey<sup>1</sup>, A. J. Macdonald<sup>1</sup>, P. R. Poulton<sup>1</sup>, T. Scott<sup>1</sup>, I. H. Köhler<sup>2†</sup>, H. Schnyder<sup>2</sup>, K. W. T. Goulding<sup>1</sup> & M. J. Crawley<sup>3</sup>

The negative effect of increasing atmospheric nitrogen (N) pollution on grassland biodiversity is now incontrovertible<sup>1–3</sup>. However, the recent introduction of cleaner technologies in the UK has led to reductions in the emissions of nitrogen oxides, with concomitant decreases in N deposition<sup>4</sup>. The degree to which grassland biodiversity can be expected to ‘bounce back’ in response to these improvements in air quality is uncertain, with a suggestion that long-term chronic N addition may lead to an alternative low biodiversity state<sup>5</sup>. Here we present evidence from the 160-year-old Park Grass Experiment at Rothamsted Research, UK<sup>6</sup>, that shows a positive response of biodiversity to reducing N addition from either atmospheric pollution or fertilizers. The proportion of legumes, species richness and diversity increased across the experiment between 1991 and 2012 as both wet and dry N deposition declined. Plots that stopped receiving inorganic N fertilizer in 1989 recovered much of the diversity that had been lost, especially if limed. There was no evidence that chronic N addition has resulted in an alternative low biodiversity state on the Park Grass plots, except where there has been extreme acidification, although it is likely that the recovery of plant communities has been facilitated by the twice-yearly mowing and removal of biomass. This may also explain why a comparable response of plant communities to reduced N inputs has yet to be observed in the wider landscape.

Total emissions of oxidized plus reduced N from intensive agriculture and the burning of fossil fuels increased markedly from the middle of the twentieth century in industrialized nations<sup>7</sup>. There is strong evidence from comparisons of similar habitats along N deposition gradients<sup>3,8</sup> that these increases have led to declining biodiversity in semi-natural ecosystems through acidification and eutrophication. These ‘space-for-time’ studies assume that air pollution has only increased, and that the deposition gradient is representative of a unidirectional temporal shift in grassland biodiversity. Since the late 1980s, however, measures to reduce atmospheric pollution have successfully reduced UK emissions of NO<sub>x</sub> by ~50% and of sulfur (S) by ~90% (ref. 9). Quantifying the potential recovery of biodiversity in response to reducing air pollution requires an alternative to the space-for-time approach, ideally monitoring long-term community dynamics on permanent plots<sup>2</sup>. In this context, the Park Grass Experiment at Rothamsted, which started in 1856, presents a unique opportunity to study shifts in biodiversity in response to environmental change both pre- and post-industrialization<sup>6</sup>.

Park Grass consists of permanent plots with different fertilizer treatments that were established on a uniform pasture that was at least 100 years old in 1856. In the early 1900s, most plots were divided in two, and lime was applied to one half—designated the limed (L) or unlimed (U) sub-plots. In 1965, the limed sub-plots were further split into sub-plots ‘a’ and ‘b’, and the unlimed sub-plots were further divided into sub-plots ‘c’ and ‘d’. Since this time, varying amounts of lime have periodically been added to maintain a target pH of 7, 6 and 5 for sub-plots a, b and c, respectively; sub-plot d is left unlimed (Extended Data Table 1). The liming treatments mean that the eutrophication effect of atmospheric N deposition on plant community dynamics can be quantified independently of soil pH (which also responds to changes in S

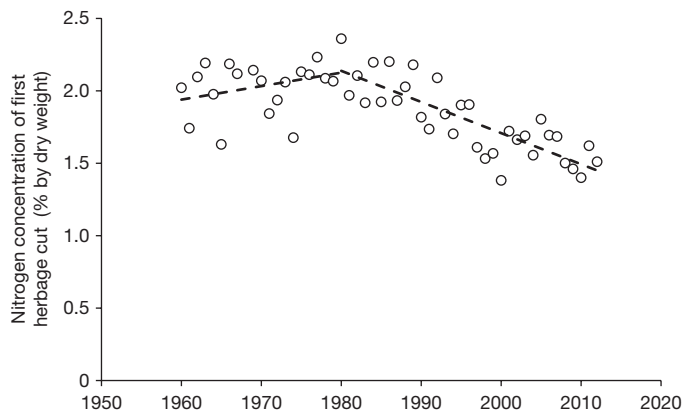
deposition). Park Grass is in a semi-urban environment, close to a road and on the edge of the town of Harpenden, which act as local sources of atmospheric pollutants<sup>4,10</sup>. Local measurements of ammonium and nitrate deposited in rainfall show that they have both declined by a comparable amount since 1985, and reflect the current national downward trend in total N emissions (Fig. 1). Our measurements did not



**Figure 1 | Changes in atmospheric N deposition, pH and proportion of legumes on the limed and unlimed sub-plots of the Park Grass nil plot.** **a**, Changes in wet (x) and dry (+) N deposition; line indicates moving 5-year average (the small increase in the early 2000s may be a legacy of a run of mild winters in the 1990s). **b**, Change in proportion of legumes (by dry weight) measured in the first herbage cut. Lines indicate the change in decadal average of percentage legumes on the limed plot (blue circles, blue dashed line), or unlimed plot (red circles, red dashed line). The plots had an average pH over the period of 7.0 when limed (blue diamond, blue continuous line) and 5.2 when unlimed (red diamond, red continuous line).

<sup>1</sup>Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK. <sup>2</sup>Lehrstuhl für Grünlandlehre, Technische Universität München, Alte Akademie 12, 85354 Freising-Weihenstephan, Germany.

<sup>3</sup>Department of Biological Sciences, Imperial College London, Silwood Park, Ascot, Berkshire SL5 7PY, UK. <sup>†</sup>Present address: Global Change and Photosynthesis Research Unit, Agricultural Research Service, United States Department of Agriculture, Urbana, Illinois 61801, USA.



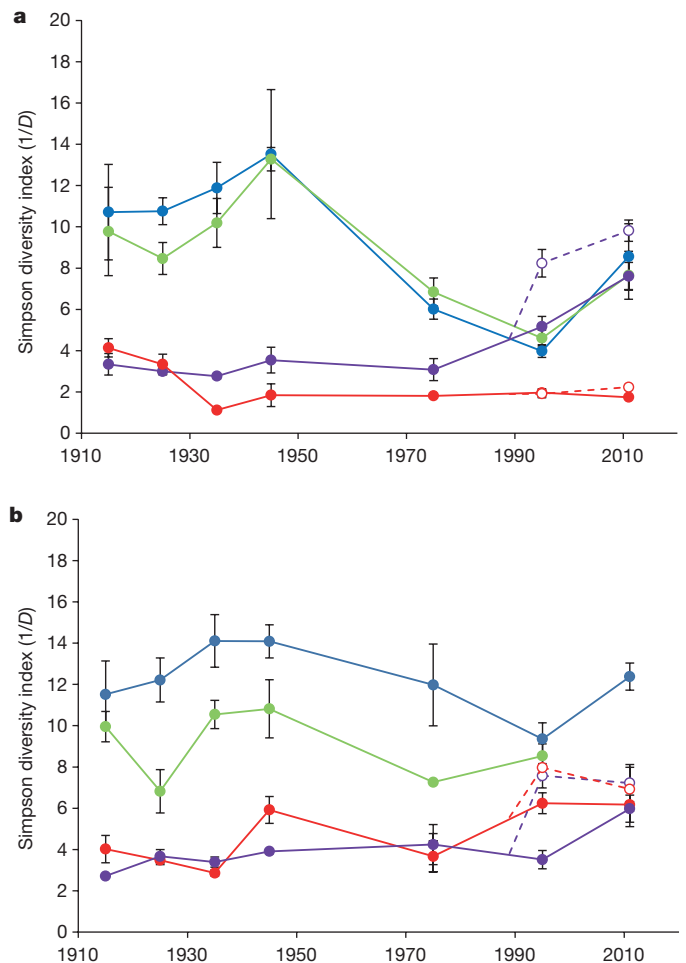
**Figure 2 | Change in N concentration measured on archived herbage samples taken from Park Grass sub-plot 3d between 1960 and 2012 that has never received any external inputs of lime or fertilizer.**

A split-line regression fitted to the data has an intersection point of 1980 (95% confidence limits, 1976 and 1988), with a significant decline after 1980 ( $R^2 = 0.67$ ,  $P < 0.001$  using least squares linear regression). During the sampling period, legumes never exceed 5% of total biomass, and the contribution of nitrogen fixation to N in the herbage is therefore expected to be minimal.

include all species of N, including ammonia; however, estimated total N deposition for grassland at this site in 2010–2012, including all N species, was  $\sim 21 \text{ kg ha}^{-1} \text{ yr}^{-1}$  (<http://www.apis.ac.uk>) compared with  $\sim 45 \text{ kg ha}^{-1} \text{ yr}^{-1}$  measured in 1996 (ref. 4). Analysis of the N concentration in archived herbage samples from the first hay cut from the plot that has never received any fertilizer inputs also showed a significant decline in percentage N since the 1980s (Fig. 2).

We analysed data on relative biomass of vascular plant species sampled on a range of sub-plots between 1903 and 2012, with a focus on both the 'nil plot' (plot 3), which has never received any fertilizers, and the 'transition plots' (Extended Data Table 2). The latter received  $96 \text{ kg N ha}^{-1}$  (plus P, K, Na and Mg), either as ammonium sulfate (plot 9) or sodium nitrate (plot 14) until 1989, when the plots were split. Since then, no further N was applied to one-half of the plots (now 9/1 and 14/1). The original treatment continued on the remaining halves (now plots 9/2 and 14/2). Generalized linear models (GLMs) and mixed models (GLMMs) were used to quantify the effect of changes in wet atmospheric N deposition (measured as either a 3- or 5-year moving average) on the proportion of plant functional groups, species richness and the exponent of the Shannon diversity index ( $e^{H'}$ ). Temporal trends in relative species abundance and dissimilarity between plots were also analysed using multivariate methods.

On the nil plot, the proportion of legumes tracked changes in atmospheric N deposition, declining to low relative abundance at the end of the twentieth century before showing a degree of recovery over the recent sampling period (Fig. 1 and Extended Data Table 3). The addition of lime also increased the proportion of legumes and forbs at the expense of grasses. A decrease in pH was observed between 1985 and 1991 resulting from the deposition of S and N that was not always compensated for by the addition of lime<sup>11</sup> (Extended Data Table 1). It is likely that this contributed to some of the observed decline in the proportion of legumes at this time. However, in the recent sampling period (1991–2012), pH has largely remained constant while N deposition has continued to decline, and pH was not significantly correlated with wet N deposition in any of the models, allowing them to be treated as independent variables. Comparisons of species richness between historical sampling periods are confounded by the fact that the area sampled and protocol used has changed through time. However, as the Simpson diversity index has a low sensitivity to sample size, it can be used as an indication of temporal trends in species diversity on Park Grass (Fig. 3). A decline in diversity was observed on the nil sub-plots between the 1940s and 1990s—declines were steeper on the unlimed



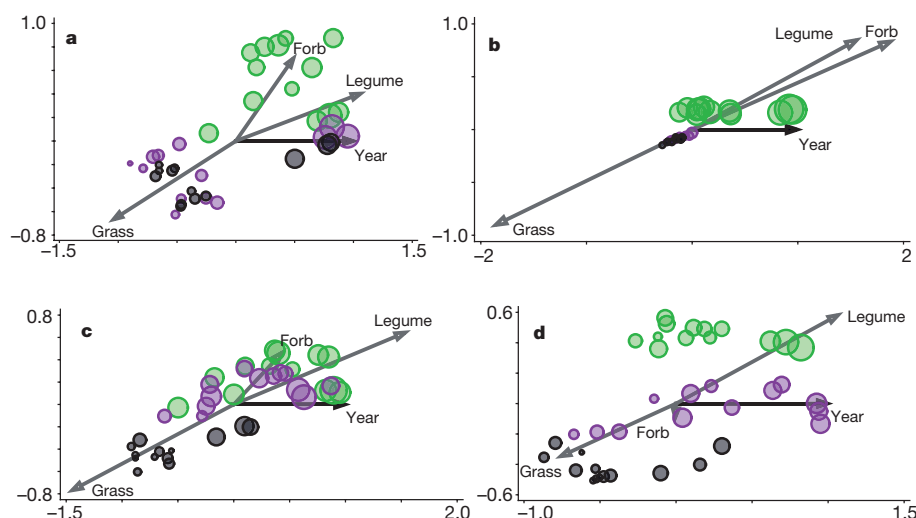
**Figure 3 | Historical trends in Simpson's plant diversity index between 1910 and 2012 for unlimed and limed sub-plots. a, b, Results are from unlimed (a) and limed (b) sub-plots.**

Decadal mean and s.e.m. are presented for: plot 3 (no fertilizers; blue circles), plot 7 (PKNaMg; green circles), plot 9/2 (PKNaMg plus  $96 \text{ kg N ha}^{-1}$  applied as ammonium sulfate; red circles), plot 14/2 (PKNaMg plus  $96 \text{ kg N ha}^{-1}$  applied as sodium nitrate; purple circles), plot 9/1 (N withheld since 1989; red open circles, dashed line) and plot 14/1 (N withheld since 1989; purple open circles, dashed line). For the limed plots in b, the data post-1965 are from the 'a' sub-plots, as they are closest to the pH maintained on the limed half of the plots before they were split. The relatively high value for plot 14/1d compared to plot 3d and 7d in the 1990s can be explained by the temporary increase in diversity during the transition period. Plot 7a was not sampled between 2010 and 2012.

sub-plot because of the combined effect of eutrophication and acidification. The latest samples, taken since 2010, show diversity is recovering, although it is still at levels below those recorded in the 1930–1940s on the unlimed nil plot (Extended Data Table 3). An analysis across all the individual sub-plots sampled confirmed the positive effect of decreasing atmospheric deposition on plant species richness and diversity, as well as an increase in the proportion of legumes (Extended Data Table 4).

The expected increase in diversity and directional shift in species communities on the transition plots after the cessation of N fertilization in 1989 was observed, except on plot 9/1d, which continues to be constrained by very low soil pH (Figs 3 and 4). Plot 7, which receives the same amount of the other nutrients as plots 9/1 and 14/1 but has never had any N fertilizer additions, can be viewed as the plot towards which the transition plots should be moving. In the case of the b sub-plots, this appears to be the case, but plot 9/1d is only recovering very slowly from a low pH, and 14/1d appears to still have a community that is intermediate between 14/2d and 7d. Over most of the recent sampling period, the plant community dynamics on plots 9/2 and





**Figure 4 | Temporal trends in plant communities, between 1991 and 2012, on selected Park Grass plots. a–d,** Redundancy analysis ordinations are presented for plot 9b (a), plot 9d (b), plot 14b (c) and plot 14d (d), including transition plots (purple circles) and plots that have continued to receive fertilizer N (grey circles). In each case, the samples from plot 7b or 7d have been included as having the species composition to which the

transition plots are moving towards (green circles). Sub-plots a and c are excluded as they were not sampled on plot 7 in 2010–2012. The size of the symbols is proportional to the numbers of species in each sample, and the relative proportion of the plant functional groups have been projected as supplementary variables.

14/2, which continued to receive N, unexpectedly showed a temporal trend that was largely parallel with the transition plots (Extended Data Figs 1 and 2 and Extended Data Tables 5 and 6). This suggests that the effect of withholding N fertilizer became apparent within the first few years of the treatment change, and since then all the plots have been responding to the same underlying environmental trend. A comparison of multivariate analyses using either year or  $\pm$ N fertilizer as the explanatory variable showed that the species that responded to withholding N on the transition plots were very similar to those driving the temporal trends in plant communities observed on the wider experiment (Extended Data Fig. 3). In particular, abundance of the legume species, *Trifolium pratense* and *Lathyrus pratensis*, and the forbs, *Plantago lanceolata* and *Ranunculus acris*, all increased significantly when N fertilizer was withheld and also across the whole experiment with time. This was confirmed from the analysis at the sub-plot level for plots 3, 9 and 14 (Extended Data Table 7).

The positive responses of plant diversity to decreasing atmospheric deposition and N fertilizer inputs on Park Grass shows that grasslands have the capacity to recover from the negative effect of eutrophication, particularly where the confounding effect of decreasing pH had been removed by applying lime. The fact that legumes showed the strongest temporal response (coinciding with measured reductions in the N concentration of the cut herbage on the nil plot) supports the view that reducing N deposition was a causal factor of the observed community dynamics. However, only wet N deposition was included in our models, and we were unable to quantify the contributions from dry deposition and other species of N over a sufficient time period, including ammonia and nitric acid. Sulfur emissions have also declined since the 1980s, and although the liming treatments mean that the indirect effect of changing S deposition on soil pH can be treated independently of the eutrophication effect of N deposition, we cannot fully discount the direct nutritional effect of S. However, the plots included in the analysis with K, Na and Mg as part of the fertilizer treatment also receive up to  $122 \text{ kg S ha}^{-1} \text{ yr}^{-1}$ , meaning they are unlikely to be limited by S.

The continuity of the experimental treatments on Park Grass, together with the measurement of atmospheric chemistry and plant community data on the same local scale, avoids some of the problems associated with attributing large-scale ecological changes observed in national vegetation surveys to anthropogenic drivers<sup>12</sup>. This may partly explain why a clear signal of a recovery of plant diversity from eutrophication has yet to be detected in the wider landscape<sup>13,14</sup>. However, it

is also the case that the magnitude of local scale reductions in N deposition we observed at Rothamsted are not yet reflected at the national scale to the same degree<sup>9</sup>. Interpreting changes on the Park Grass Experiment more widely in the context of comparisons with other grassland studies and its relevance to the wider landscape must also take into account the specific management context. The twice-yearly mowing and removal of biomass on Park Grass may explain the relatively rapid transient dynamics observed on the experiment when compared to equivalent studies in systems with less disturbance, leading to the accumulation of litter<sup>5</sup>, or dominated by slower growing, woody species<sup>15,16</sup>. In addition, the close proximity of plots with differing plant communities means that limitation of propagules is not likely to be as important a constraint in the recovery of the communities as may be the case for larger scale grassland restoration<sup>17</sup>. Despite these considerations, the Park Grass Experiment remains a unique indicator of the effects of environmental change and an important part of the evidence base for assessing the biological effects of changes in management or policy on the wider environment.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 30 June; accepted 11 November 2015.**

**Published online 2 December 2015.**

- Clark, C. M. & Tilman, D. Loss of plant species after chronic low-level nitrogen deposition to prairie grasslands. *Nature* **451**, 712–715 (2008).
- Dupră, C. *et al.* Changes in species richness and composition in European acidic grasslands over the past 70 years: the contribution of cumulative atmospheric nitrogen deposition. *Glob. Change Biol.* **16**, 344–357 (2010).
- Stevens, C. J. *et al.* Nitrogen deposition threatens species richness of grasslands across Europe. *Environ. Pollut.* **158**, 2940–2945 (2010).
- Goulding, K. W. T. *et al.* Nitrogen deposition and its contribution to nitrogen cycling and associated soil processes. *New Phytol.* **139**, 49–58 (1998).
- Isbell, F., Tilman, D., Polasky, S., Binder, S. & Hawthorne, P. Low biodiversity state persists two decades after cessation of nutrient enrichment. *Ecol. Lett.* **16**, 454–460 (2013).
- Silvertown, J. *et al.* The Park Grass Experiment 1856–2006: Its contribution to ecology. *J. Ecol.* **94**, 801–814 (2006).
- Fowler, D. *et al.* A chronology of nitrogen deposition in the UK between 1900 and 2000. *Water Air Soil Pollut. Focus* **4**, 9–23 (2004).
- Stevens, C. J., Dise, N. B., Mountford, J. O. & Gowing, D. J. Impact of nitrogen deposition on the species richness of grasslands. *Science* **303**, 1876–1879 (2004).

9. RoTAP. *Review of Transboundary Air Pollution: Acidification, Eutrophication, Ground Level Ozone and Heavy Metals in the UK*. Contract Report to the UK Government; <http://www.rotap.ceh.ac.uk/> (Centre for Ecology and Hydrology, 2012).
10. Zhao, F. J., Knights, J. S., Hu, Z. Y. & McGrath, S. P. Stable sulfur isotope ratio indicates long-term changes in sulfur deposition in the broadbalk experiment since 1845. *J. Environ. Qual.* **32**, 33–39 (2003).
11. Blake, L., Goulding, K. W. T., Mott, C. J. B. & Johnston, A. E. Changes in soil chemistry accompanying acidification over more than 100 years under woodland and grass at Rothamsted Experimental Station, UK. *Eur. J. Soil Sci.* **50**, 401–412 (1999).
12. Smart, S. M. *et al.* Clarity or confusion? Problems in attributing large-scale ecological changes to anthropogenic drivers. *Ecol. Indic.* **20**, 51–56 (2012).
13. Maskell, L. C., Smart, S. M., Bullock, J. M., Thompson, K. & Stevens, C. J. Nitrogen deposition causes widespread loss of species richness in British habitats. *Glob. Change Biol.* **16**, 671–679 (2010).
14. van den Berg, L. J. L. *et al.* Direct and indirect effects of nitrogen deposition on species composition change in calcareous grasslands. *Glob. Change Biol.* **17**, 1871–1883 (2011).
15. Power, S. A., Green, E. R., Barker, C. G., Bell, J. N. B. & Ashmore, M. R. Ecosystem recovery: heathland response to a reduction in nitrogen deposition. *Glob. Change Biol.* **12**, 1241–1252 (2006).
16. Terry, A. C., Ashmore, M. R., Power, S. A., Allchin, E. A. & Heil, G. W. Modelling the impacts of atmospheric nitrogen deposition on Calluna-dominated ecosystems in the UK. *J. Appl. Ecol.* **41**, 897–909 (2004).
17. Clark, C. M. & Tilman, D. Recovery of plant diversity following N cessation: effects of recruitment, litter, and elevated N cycling. *Ecology* **91**, 3620–3630 (2010).

**Acknowledgements** We thank the large teams of people who were involved in the vegetation sampling and sorting between 1991 and 2012, and J. Lepš and P. Šmilauer for their advice on the multivariate analysis. Park Grass is supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC) and the Lawes Agricultural Trust. The wet N deposition (precipitation chemistry) data set for 1992–2013 was provided courtesy of the UK Environmental Change Network (ECN). I.H.K. was supported by Deutsche Forschungsgemeinschaft (DFG SCHN 557/5-1).

**Author Contributions** J.S., M.J.C. A.J.M., P.R.P. and T.S. co-ordinated and contributed to the vegetation sampling between 1991 and 2012. T.S. and K.W.T.G. were responsible for collecting and analysing nitrogen deposition data. I.H.K. and H.S. analysed nitrogen limitation of vegetation. J.S. was responsible for the statistical analysis and initial draft of the paper. All authors contributed to the final version.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.S. ([jonathan.storkey@rothamsted.ac.uk](mailto:jonathan.storkey@rothamsted.ac.uk)).

## METHODS

**Description of Park Grass Experiment and vegetation sampling protocol.** The Park Grass Experiment was established on old grassland at Rothamsted in 1856 to examine the effects of different mineral fertilizers and organic manures on productivity of permanent pasture cut for hay<sup>6</sup>. The experiment is located on a moderately well-drained silty clay loam overlying clay-with-flints, a chronic or vertic Luvisol according to the FAO classification. The soil pH was slightly acidic when the experiment began (5.4–5.6), and the nutrient status was poor. The original vegetation of Park Grass was classified as dicotyledon-rich *Cynosurus cristatus*–*Centaurea nigra* grassland; one of the mesotrophic grassland communities in the British National Vegetation Classification system<sup>18</sup>. Treatments imposed in 1856–1865 included controls (nil, no fertilizer or manure), and various combinations of P, K, S, Mg and Na, with N applied as either sodium nitrate or ammonium salts. Each plot (ranging from 75 to 634 m<sup>2</sup>) now consists of plant communities adapted to the fertilizer treatments naturally assembled from the local species pool. Farmyard manure was applied to two plots but was discontinued after 8 years, because when applied annually to the surface in large amounts it did not decompose quickly and had adverse effects on the sward. Farmyard manure, applied every 4 years, was re-introduced on three plots in 1905.

The experiment consists of 20 main plots. The plots are cut in mid-June and made into hay. For 19 years, the re-growth was grazed by sheep penned on individual plots, but since 1875 a second cut, usually carted green, has been taken in place of grazing. The plots were originally cut by scythe, then by horse-drawn and then tractor-drawn mowers. Yields were originally estimated by weighing the produce, either of hay (first harvest) or green crop (second harvest), and dry matter was determined from the whole plot. Since 1960, yields of dry matter have been estimated from strips cut with a forage harvester. However, for the first cut the remainder of the plot is still mown and made into hay, continuing earlier management and ensuring the return of seed. For the second cut, the whole plot is cut with a forage harvester. A small amount of lime, 4 t CaCO<sub>3</sub> ha<sup>-1</sup>, was added to all plots in the late 1880s. Most plots were divided in two in 1903 or 1920 to introduce a test of regular liming on one half. In 1965, they were further divided into four sub-plots (a–d). The a, b and c sub-plots now receive lime every 3 years, if necessary, sufficient to maintain a target soil pH of 7, 6 and 5, respectively. The d sub-plots are unlimed. In 1989, the plots receiving 96 kg N ha<sup>-1</sup> were split, and nitrogen fertilizer withheld from half of the plots to investigate the ability of the plant communities to recover from chronic nitrogen addition.

Vegetation surveys have been carried out on Park Grass on more than 30 occasions since the experiment began<sup>19</sup>. The original botanical sampling protocol was to take handfuls of cut herbage at regular intervals from every swath after the scythe or cutting machine. Each sample was then sub-sampled until a weight of approximately 12–20 lb (5.4–9.1 kg) was obtained. For the samples taken between 1973 and 1976, samples were cut by hand every two to three paces along ten transects on the larger plots and six on the smaller plots. Approximately 600 g of material was analysed from each sub-plot. Between 1991 and 2012, above-ground biomass from six randomly located 50 × 25 cm quadrats was sampled from all sub-plots using a standard protocol. The herbage was cut with scissors to ground level in early June, immediately before harvesting the first hay crop. The plant material was taken back to the laboratory where it was sorted into species. Samples were oven-dried at 80 °C for about 24 h, after which dry mass was determined for each species. Data from the six quadrats were aggregated to provide an estimate of species richness for each plot in each year.

**Monitoring of atmospheric nitrogen deposition: wet ‘bulk’ deposition.** The methods of collection and the amount of data available have varied over time. Initial precipitation data (1853–1968) were collected using a rain gauge with a surface area of one-thousandth of an acre (7 ft 3.12 in. × 6 ft, or 4.04 m<sup>2</sup>). The gauge being constructed at ground level of lead supported by wood over a brick lined cellar housing four collection tanks from which a sample of rain water was taken. From 1969 to 1986, precipitation was collected in what is described as a ‘simple funnel-and-bottle bulk gauge’<sup>20</sup>. In the latter years, 1986 to present, precipitation has been collected in a bulk rain water collector of a design described previously<sup>21</sup>. All these collection methods took place within the Rothamsted meteorological enclosure, which is located approximately 817 m east-northeast of the Park Grass Experiment. The amounts of nitrate-N and ammonium-N, in mg l<sup>-1</sup> in solution, were then determined. The amount of nitrogen in kg ha<sup>-1</sup> deposited by wet deposition is determined by the formula: kg ha<sup>-1</sup> = (mg l<sup>-1</sup> × R)/(A × 10<sup>5</sup>), in which R is the amount of rain water collected in mm, and A is the surface area of collector funnel in m<sup>2</sup>. The total amount of nitrogen was then calculated for each year.

**Monitoring of atmospheric nitrogen deposition: dry deposition.** Nitrogen deposition in the form of NO<sub>2</sub> was collected passively using diffusion samplers over an exposure period of 2 weeks. The samplers are made up of a 30-μl aliquot of 20% triethanolamine/water absorbent sandwiched between two stainless steel meshes housed in a coloured thermoplastic rubber cap. Into this

cap, a 70-mm-long × 11-mm-diameter acrylic tube is inserted with a protective white thermoplastic rubber cap on the opposite end. These samplers were then placed, in sets of three, at locations around the edge of the Park Grass Experiment at a height of 1.5 m above ground and the protective cap removed. After 2 weeks, samplers were sealed using a protective cap and collected. They were then extracted into 2 ml 18.2 MΩ RO (reverse osmosis) water and analysed for nitrite N (NO<sub>2</sub>-N) by continuous colourimetric flow analysis. The resulting levels of NO<sub>2</sub>-N in μg N m<sup>-3</sup> were then averaged over the year and converted to kg N ha<sup>-1</sup> by multiplying by the deposition velocity for managed grassland at Rothamsted (0.3751 mm s<sup>-1</sup>)<sup>22</sup>.

**Analysis of herbage samples for nitrogen concentration.** Representative sub-samples of plant material from the archived hay or herbage samples on sub-plot 3d were dried at 40 °C for 48 h, ball-milled to a homogenous fine powder, dried again at 60 °C for 24 h and analysed with an elemental analyser (NA 1110; Carlo Erba) interfaced (ConFlo II, Finnigan MAT) to a continuous-flow isotope ratio mass spectrometer (Delta Plus; Finnigan MAT), EA-CF-IRMS. After every tenth sample, a solid internal laboratory standard (SILS) with similar C/N ratio as the respective sample material (fine ground wheat flour) was run as a control. The precision (s.d.) for sample repeats was better than 0.04%. Samples taken from herbage cut between 1960 and 2012 were analysed, before this date, the herbage sampling protocol differed with material dried *in situ*, which is affected by disintegration losses in the hay making process.

**Statistics.** For the plots where all the sub-plots were sampled (3, 9 and 14), all sub-sets regression using GLMs was used to identify the model that explained the maximum variability in species richness, *e*<sup>HT</sup> and the relative proportion of functional groups using only independent explanatory variables with *P* < 0.05. The following explanatory variables were included: pH, wet atmospheric nitrogen deposition (included both as a 3- and 5-year moving average), total rainfall in the previous growing season (March–August) and rainfall in current growing season (March–May)—rainfall has been found to explain short-term variability in community composition significantly<sup>23</sup>. For the proportion of the different functional groups (legumes, grasses and ‘other’), a binomial distribution with a logit link function was used to allow for the variability in the total first cut biomass to be accounted for. The proportion of each functional group was analysed separately. A normal distribution with an identity link was used for species richness and *e*<sup>HT</sup> except for the acid plots with a high frequency of low species counts, in which a Poisson distribution with a log link was used. As opposed to a step-wise approach, all sub-sets regression analyses included all possible combinations of explanatory variables, using the adjusted R<sup>2</sup> and Mallows’ C<sub>p</sub> as criteria for comparing models.

For the nil treatment (plot 3), that has never received any fertilizers, data were available on relative proportion of functional groups that covered an important historical period from 1903 in which wet nitrogen deposition and pH were also measured on the experiment. This whole data set (*n* = 70) was therefore included in the models. Although less frequent data were available on species richness and diversity, changes in the area sampled and sorting effort meant that a comparison of data over the whole historical period would not have been valid. However, between 1991 and 2012, a standard area and sampling protocol was used. This time period coincided with the reductions in atmospheric nitrogen deposition observed on the experiment and was, therefore, used to quantify responses of species richness and *e*<sup>HT</sup> (*n* = 13) across all the plots.

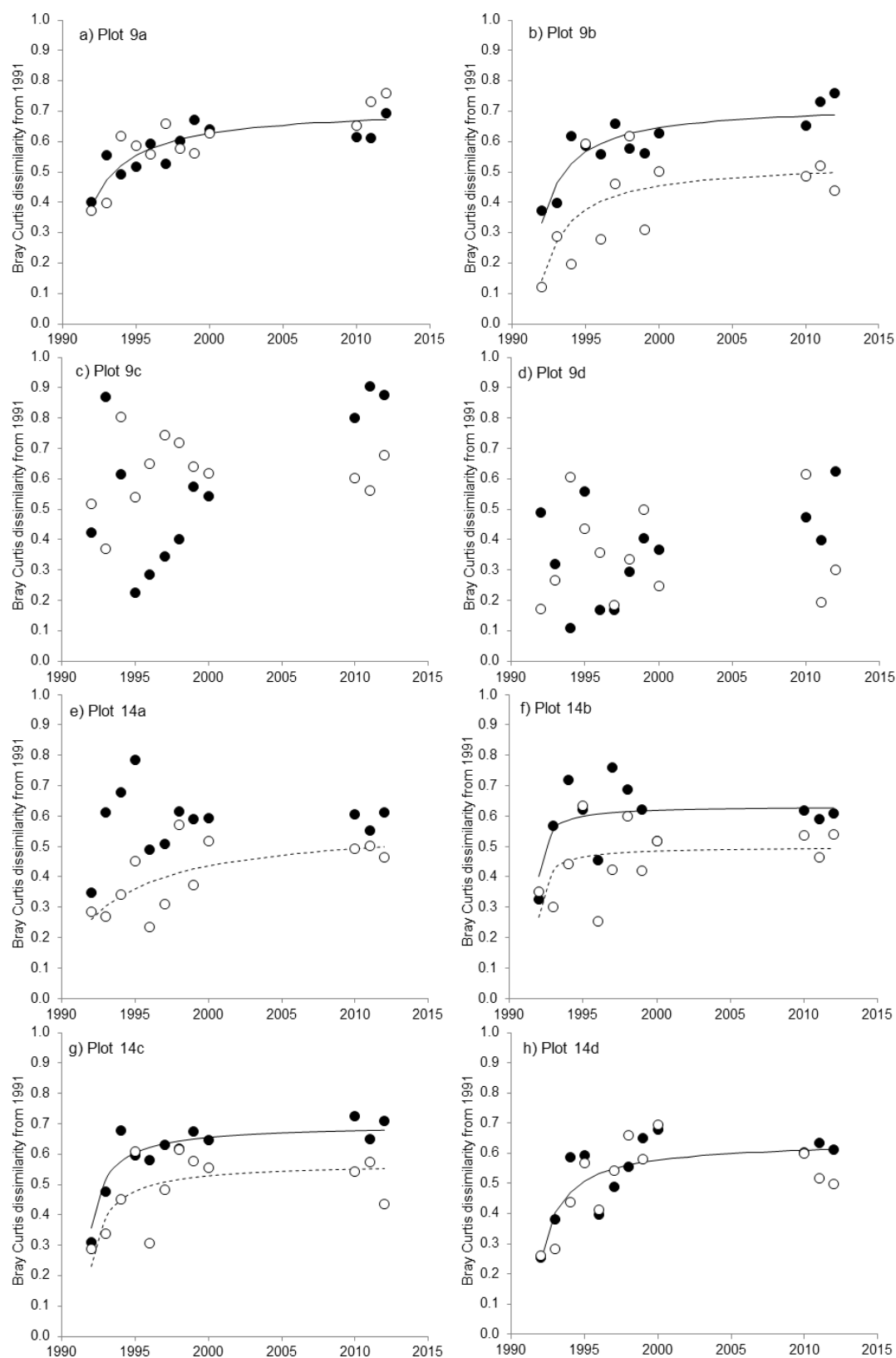
For the analysis of the data from all sub-plots sampled across a range of fertilizer treatments, a GLMM was used with sub-plot and year input as random factors. Fixed effects were input in the same order to a model previously fitted to explain variance in species richness between fertilizer treatments<sup>24</sup>: pH, nitrogen addition (three levels; +48, +96 or +144 kg N ha<sup>-1</sup>) and ±phosphorus before inputting N deposition as a final continuous explanatory variable as either a 3- or a 5-year moving average. Two further fixed effects were initially included, ±potassium and whether the plot was a transition plot, but neither significantly explained any additional variance in any of the diversity metrics. Plot 13 was the only farmyard manure plot in the data set and was not included in this analysis. Both the GLMs and the GLMMs were run using the software GenStat<sup>25</sup>.

The temporal shifts in plant communities was analysed at the species level using multivariate approaches. To investigate any directional response of the transition plots following the cessation of nitrogen fertilization, the Bray–Curtis dissimilarity index was calculated using the first sample date, 1991, as a reference point and regression models fitted to the data using year as the explanatory variable. This was also done for the plots that continued to receive nitrogen fertilizer with the expectation that these plots would increasingly diverge from the transition plots with time. Nonlinear regression with groups was used to quantify differences in the responses of the Bray–Curtis index to time of the transition plots and those that had continued to receive nitrogen. Redundancy analysis, using rainfall in the current growing season as a covariate, was used to identify community shifts over time, using GLMs to identify species that responded significantly to the first ordination axis,



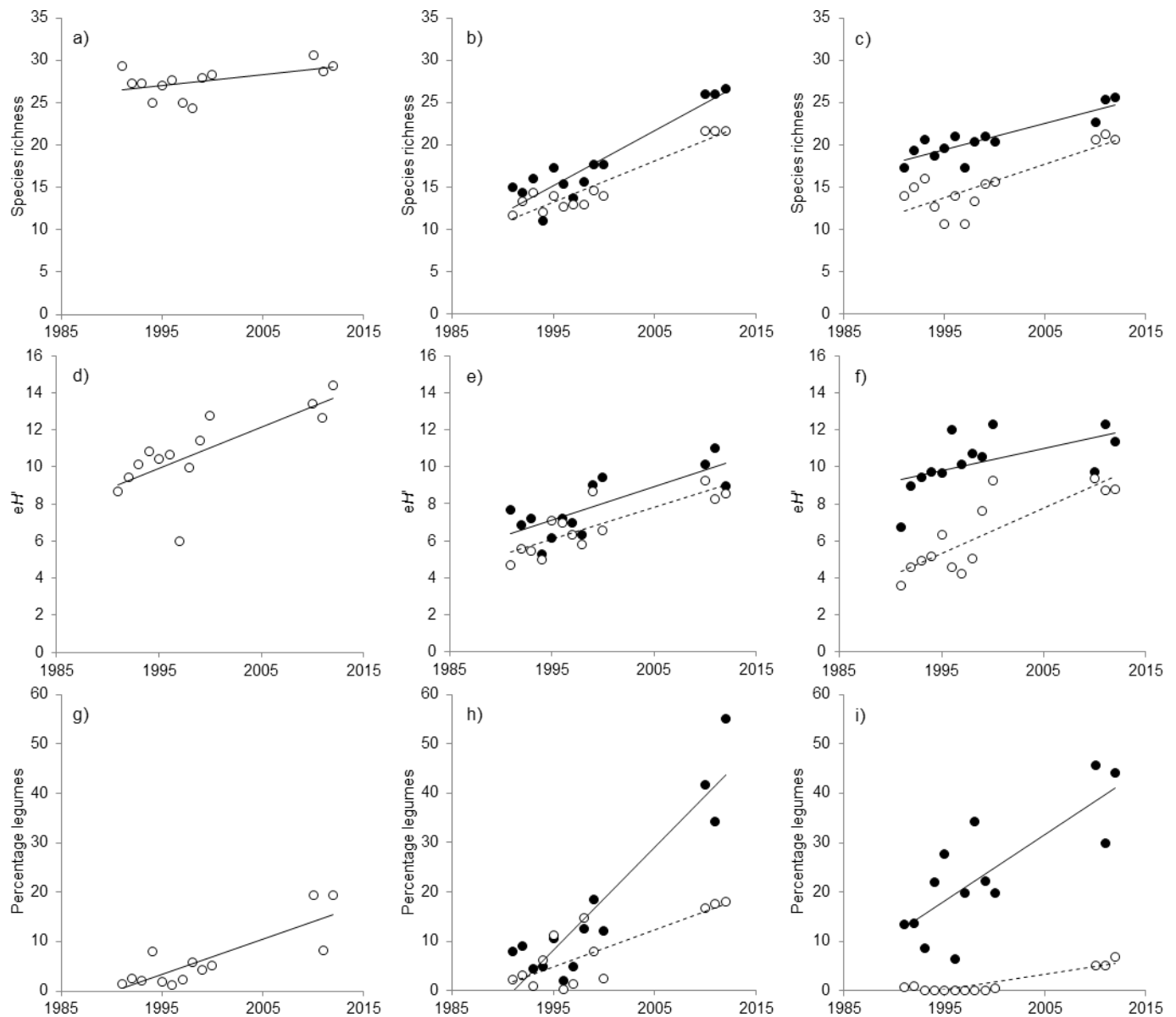
constrained by year. This was done for each of the sub-plots separately for plots 3, 9/1, 9/2, 14/1 and 14/2. Finally, the species responding to withholding N fertilizer were compared with those driving the temporal responses using two additional partial canonical correspondence analyses. First, data from 1991–2012 for plots 9 and 14 were analysed with year input as a categorical covariate and  $\pm$ nitrogen as the explanatory variable. Second, the data from all the sub-plots sampled from 1991–2012 were analysed, excluding the transition plots (9/1 and 14/1) with plot input as a covariate, and year as a continuous explanatory variable. Only species that were recorded at least three times at the level of the sub-plot were included in the analysis. The software, Canoco 5, was used for all the multivariate analyses<sup>26</sup>.

18. Dodd, M. E., Silvertown, J., McConway, K., Potts, J. & Crawley, M. Application of the British National Vegetation Classification to the communities of the Park Grass Experiment through time. *Folia Geobot. Phytotaxon.* **29**, 321–334 (1994).
19. Williams, E. D. *Botanical Composition of the Park Grass Plots*; Rothamsted Experimental Station Report for 1977, part 2, 31–36 (Lawes Agricultural Trust, 1978).
20. Goulding, K. W. T., Poulton, P. R., Thomas, V. H. & Williams, R. J. B. Atmospheric deposition at Rothamsted Experimental Station, Saxmundham Experimental Station and Woburn Experimental Station, England, 1969–1984. *Wat. Air Soil Pollut.* **29**, 27–49 (1986).
21. Hall, D. J. *The Precipitation Collector for use in the Secondary National Acid Deposition Network* (Warren Spring Laboratory, 1986).
22. Smith, R. I., Fowler, D., Sutton, M. A., Flechard, C. & Coyle, M. Regional estimation of pollutant gas dry deposition in the UK: model description, sensitivity analyses and outputs. *Atmos. Environ.* **34**, 3757–3777 (2000).
23. Silvertown, J., Dodd, M. E., McConway, K., Potts, J. & Crawley, M. Rainfall, biomass variation, and community composition in the Park Grass Experiment. *Ecology* **75**, 2430–2437 (1994).
24. Crawley, M. J. *et al.* Determinants of species richness in the park grass experiment. *Am. Nat.* **165**, 179–192 (2005).
25. Payne, R. W., Murray, D. A. & Harding, S. A. *An Introduction to the GenStat Command Language* 14 edn (VSN International, 2011).
26. Smilauer, P. & Lepš, J. *Multivariate Analysis of Ecological Data Using Canoco 5* (Cambridge Univ. Press, 2014).



**Extended Data Figure 1 | The Bray–Curtis dissimilarity index. a–h,** The response of Bray–Curtis dissimilarity for all sub-plots on plot 9 (a–d) and plot 14 (e–h). Community data from 1992–2012 have been compared to samples taken in 1991 for the transition plots (filled circles) and plots that

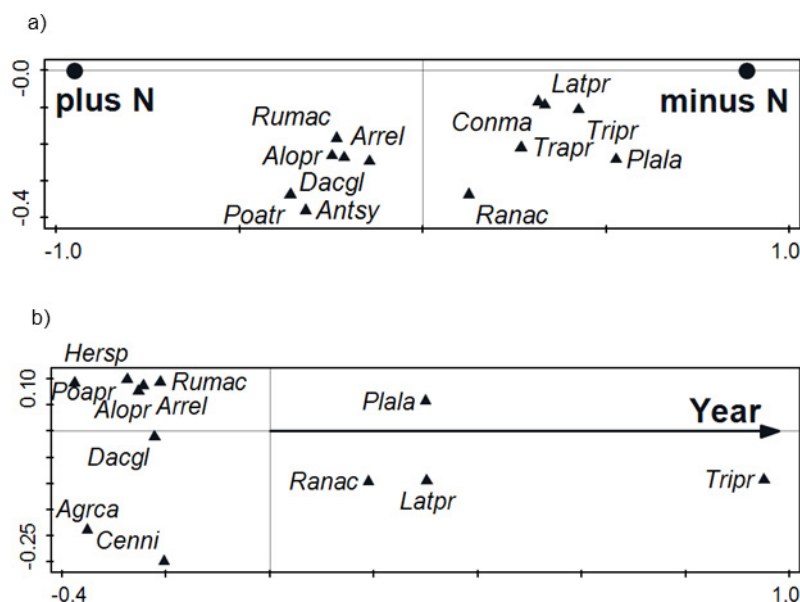
continue to receive inorganic N fertilizer (open circles). Lines indicate a significant fit for a rectangular hyperbola function; where separate lines have been fitted in a single panel, a significant difference in the asymptote of the responses was observed.



**Extended Data Figure 2 | Change in time over the recent sampling period (1991–2012) of species richness,  $eH'$ , and percentage legumes in the first herbage cut. a–i, Plot 3 (a, d, g), plot 9 (filled circles, 9/1; open circles, 9/2) (b, e, h) and plot 14 (filled circles, 14/1; open circles, 14/2)**

(c, f, i). The averages of sub-plots a, b and c are presented; sub-plot d was excluded to avoid the confounding effect of very low pH on plot 9/1d allowing a direct comparison between treatments at the main plot level.





**Extended Data Figure 3 | Comparison of the effect of decreasing atmospheric N inputs and the cessation of N fertilization on plant communities.** **a**, Partial canonical correspondence analysis (CCA) of the effect of withholding nitrogen fertilizer on plant communities on the transition plots 9/1 and 14/1 compared to the plots that continued to receive N, 9/2 and 14/2. Data from all sub-plots during the modern day sampling period (1991–2012) were used, and year was included as a categorical covariate. **b**, Partial CCA of the temporal response of plant communities on all sub-plots sampled during the modern day period, 1991–2012, excluding the transition plots 9/1 and 14/1, with year entered

as a continuous variable and plot as a covariate. In both ordination plots, species were only included if they were in the top 20 species ranked by their weighting in the CCA and had a *P* value indicating their association with the constrained axis of  $<0.1$ . *Agrca*, *Agrostis capillaris*; *Alopr*, *Alopecurus pratensis*; *Antsy*, *Anthriscus sylvestris*; *Arrel*, *Arrhenatherum elatius*; *Conma*, *Conopodium majus*; *Dacgl*, *Dactylis glomerata*; *Hersp*, *Heracleum sphondylium*; *Latpr*, *Lathyrus pratensis*; *Plala*, *Plantago lanceolata*; *Poapr*, *Poa pratensis*; *Poatr*, *Poa trivialis*; *Ranac*, *Racunculus acris*; *Rumac*, *Rumex acetosa*; *Trapr*, *Tragopogon pratense*; *Tripr*, *Trifolium pratense*.

**Extended Data Table 1 | pH (in water) measured on soil cores, 0–23 cm, on seven occasions between 1991 and 2012**

Plot	Treatment	Nov-91	Feb-95	Mar-98	Mar-02	Mar-05	Mar-08	Mar-11
1b	N <sub>1</sub>	6.4	7.0	7.1	7.3	7.2	7.1	6.9
1d		3.6	4.2	4.0	4.1	4.0	3.8	4.0
2/2b	Nil	6.5	6.9	6.9	7.4	7.5	7.3	7.1
2/2b		4.8	5.0	5.2	5.1	5.2	5.1	5.1
3a	Nil	6.4	7.1	7.1	7.3	7.3	7.2	7.2
3b		6.4	6.5	6.5	6.3	6.4	6.1	6.3
3c		5.0	5.4	5.3	5.2	5.1	4.9	5.2
3d		4.8	5.2	5.2	5.3	5.3	5.2	5.3
7b	PKNaMg	6.2	6.0	5.9	5.8	6.0	6.1	6.2
7d		4.8	5.0	4.7	4.9	4.9	4.9	4.9
9/1a	(N <sub>2</sub> )PKNaMg	5.7	6.5	6.5	6.9	7.0	7.1	7.1
9/1b		5.2	5.9	6.1	6.4	6.1	6.1	6.4
9/1c		4.4	4.6	4.8	5.3	5.0	4.9	5.2
9/1d		3.7	4.1	4.0	4.1	4.0	4.0	4.1
9/2a	N <sub>2</sub> PKNaMg	6.2	6.8	6.8	7.1	7.2	7.1	7.1
9/2b		5.4	6.2	6.2	6.4	6.0	6.3	6.2
9/2c		4.4	5.3	4.8	4.8	5.4	4.8	5.1
9/2d		3.6	3.9	3.6	3.6	3.6	3.4	3.7
10b	N <sub>2</sub> PNaMg	5.4	5.5	5.9	6.4	6.7	6.2	6.3
10d		3.4	3.8	3.7	3.7	3.7	3.5	3.7
11/1b	N <sub>3</sub> PKNaMg	5.4	6.4	6.1	6.2	6.4	6.0	6.4
11/1d		3.2	3.7	3.5	3.6	3.5	3.4	3.6
11/2b	N <sub>3</sub> PKNaMgSi	5.2	5.9	5.6	6.3	6.5	6.1	6.1
11/2d		3.4	3.9	3.7	3.6	3.6	3.4	3.6
13/2b	FYM / PM	6.3	6.1	6.2	5.9	6.0	5.9	6.1
13/2d		4.6	5.1	5.1	5.1	5.2	5.2	5.0
14/1a	(N* <sub>2</sub> )PKNaMg	6.6	6.9	6.8	7.0	6.9	7.0	6.9
14/1b		6.6	6.3	6.1	6.0	5.7	6.2	6.0
14/1c		5.8	5.8	5.6	5.5	5.4	5.3	5.3
14/1d		5.6	5.8	5.6	5.6	5.6	5.6	5.4
14/2a	N* <sub>2</sub> PKNaMg	6.8	6.8	6.9	7.0	7.0	7.0	7.0
14/2b		6.8	6.4	6.5	6.3	6.3	6.3	6.2
14/2c		6.0	6.1	6.1	6.0	6.1	6.0	5.9
14/2d		5.8	5.7	6.1	6.0	6.1	6.1	6.0
17b	N* <sub>1</sub>	6.7	6.5	6.6	6.2	6.4	6.2	6.3
17d		5.6	5.8	5.8	5.7	6.0	5.8	5.7

Treatments: N<sub>1</sub>, N<sub>2</sub> and N<sub>3</sub>: ammonium sulfate supplying 48, 96 and 144 kg N ha<sup>-1</sup>; 55, 110 and 165 kg S ha<sup>-1</sup>; N\*<sub>1</sub> and N\*<sub>2</sub>: sodium nitrate supplying 48 and 96 kg N and 78 and 157 kg Na ha<sup>-1</sup>; (N<sub>2</sub>), (N\*<sub>2</sub>): N last applied 1989; P: triple superphosphate supplying 35 kg P ha<sup>-1</sup>; K: potassium sulfate supplying 225 kg K and 99 kg S ha<sup>-1</sup>; Na: sodium nitrate supplying 15 kg Na and 10 kg Na ha<sup>-1</sup>; Mg: magnesium sulfate (Epsom salts) supplying 10 kg Mg and 13 kg S ha<sup>-1</sup>; Si: water-soluble sodium silicate supplying 135 kg Si and 63 kg Na ha<sup>-1</sup>; FYM: 35 t farmyard manure ha<sup>-1</sup> supplying ~240 kg N, 45 kg P, 350 kg K, 25 kg Na, 25 kg Mg, 40 kg S, 135 kg Ca ha<sup>-1</sup>; PM: pelleted poultry manure (replaced fishmeal in 2003) supplying ~65 kg N ha<sup>-1</sup>. Sub-plots a, b and c receive differential amounts of lime, if needed, every 3 years to maintain soil pH at 7, 6 and 5, respectively; sub-plot d receives no lime.

**Extended Data Table 2 | Species richness and  $eH'$  observed in a total sample area of 0.75 m<sup>2</sup> averaged over three biomass samples taken in 1991–1993 and 2010–2012**

Plot	Treatment	Species number (s.e.m.)		$eH'$ (s.e.m.)	
		1991-93	2010-12	1991-93	2010-12
1b	N <sub>1</sub>	22(1.5)	22(0.3)	8.4(0.42)	8.8(1.13)
1d		4(0.6)	4(0.3)	1.6(0.23)	2.3(0.19)
2/2b	Nil	27(0.9)	33(3.2)	12.4(0.48)	15(1.19)
2/2d		24(1.2)	26(0.6)	5.8(0.64)	9.9(0.7)
3a	Nil	30(1.2)	31(1.2)	11.4(0.58)	15.7(0.33)
3b		29(0)	30(1.7)	10.6(0.85)	15.1(0.22)
3c		25(3.1)	28(1.2)	6.2(0.42)	9.7(1.21)
3d		28(2.7)	27(0.9)	7.3(0.43)	11.2(1.5)
7b	PKNaMg	22(0.9)	25(1.7)	10.3(0.58)	8.9(1.31)
7d		20(0.3)	26(1.5)	7.9(0.38)	10.5(0.91)
9/1a	(N <sub>2</sub> )PKNaMg	20(1.3)	26(0.9)	11(0.46)	11.1(0.51)
9/1b		15(1.2)	26(0.3)	8.1(0.69)	7.3(0.79)
9/1c		10(0.8)	27(0.3)	2.6(0.66)	11.7(0.79)
9/1d		3(0.3)	6(1.2)	1.5(0.07)	2.4(0.17)
9/2a	N <sub>2</sub> PKNaMg	16(1.5)	21(0.6)	7(0.22)	9.3(0.64)
9/2b		13(0.9)	22(0.3)	4.8(0.24)	8.4(0.46)
9/2c		10(0.9)	22(0.3)	3.9(0.76)	8.4(0.39)
9/2d		3(0.6)	3(0)	1.5(0.2)	1.9(0.03)
10b	N <sub>2</sub> PNaMg	9(0.7)	15(1)	5(0.86)	5.5(0.37)
10d		2(0.3)	3(0.3)	1.1(0.05)	1.5(0.4)
11/1b	N <sub>3</sub> PKNaMg	12(1)	14(0.3)	4.2(0.4)	6.5(0.47)
11/1d		1(0.3)	1(0)	1(0.01)	1(0)
11/2b	N <sub>3</sub> PKNaMgSi	8(0)	14(0.6)	3.9(0.26)	4.8(0.24)
11/2d		1(0.3)	3(0.3)	1(0)	1.3(0.11)
13/2b	FYM / PM	21(0.3)	25(1.5)	9.7(0.76)	11.6(0.59)
13/2d		20(1.2)	25(0.9)	7.7(0.34)	10.7(0.51)
14/1a	(N <sup>*</sup> <sub>2</sub> )PKNaMg	20(1.2)	25(0.6)	8.7(0.38)	10(0.89)
14/1b		20(1.5)	23(1.5)	8.7(1.34)	10.8(0.78)
14/1c		18(1)	26(1.5)	7.8(0.84)	12.6(1.46)
14/1d		18(0.9)	22(0.6)	8.6(0.08)	12.4(0.33)
14/2a	N <sup>*</sup> <sub>2</sub> PKNaMg	17(1.2)	22(1.2)	4.3(0.26)	9(0.43)
14/2b		14(1.5)	20(0.6)	4.6(0.42)	8.5(0.62)
14/2c		14(1.7)	20(1.2)	4.2(0.6)	9.4(0.08)
14/2d		16(0.7)	19(0.9)	5.6(0.52)	9.8(0.31)
17b	N <sup>*</sup> <sub>1</sub>	23(3)	27(2.4)	8.2(1.28)	10.8(1.66)
17d		25(1.9)	29(0.9)	10.1(0.52)	10.8(0.46)

Sub-plots a and b have been receiving lime since the early 1900s while lime has only been added to sub-plot c since 1965, which also shows the largest increases in species richness over the modern day sampling period, possibly reflecting a continuing slow recovery from low soil pH. Treatments: N<sub>1</sub>, N<sub>2</sub> and N<sub>3</sub>: ammonium sulfate supplying 48, 96 and 144 kg N and 55, 110 and 165 kg S ha<sup>-1</sup>; N<sup>\*</sup><sub>1</sub>, N<sup>\*</sup><sub>2</sub>: sodium nitrate supplying 48 and 96 kg N and 78 and 157 kg Na ha<sup>-1</sup>; (N<sub>2</sub>), (N<sup>\*</sup><sub>2</sub>): N last applied 1989; P: triple superphosphate supplying 35 kg P ha<sup>-1</sup>; K: potassium sulfate supplying 225 kg K and 99 kg S ha<sup>-1</sup>; Na: sodium nitrate supplying 15 kg Na and 10 kg S ha<sup>-1</sup>; Mg: magnesium sulfate (Epsom salts) supplying 10 kg Mg and 13 kg S ha<sup>-1</sup>; Si: water-soluble sodium silicate supplying 135 kg Si and 63 kg Na ha<sup>-1</sup>; FYM: 35 t farmyard manure ha<sup>-1</sup> supplying ~240 kg N, 45 kg P, 350 kg K, 25 kg Na, 25 kg Mg, 40 kg S, 135 kg Ca ha<sup>-1</sup>; PM: pelleted poultry manure (replaced fishmeal in 2003) supplying ~65 kg N ha<sup>-1</sup>. Sub-plots a, b and c receive differential amounts of lime, if needed, every 3 years to maintain soil pH 7, 6 and 5 respectively; sub-plot d receives no lime.



**Extended Data Table 3 | GLMs fitted to relative biomass data from the Park Grass nil plot, which has never received any external inputs of lime or fertilizer**

Response variable	Explanatory variable	Estimate (s.e.)	<i>t</i> statistic	Degrees of freedom	<i>F</i> probability
Species number <sub>1991-2012</sub>	Soil pH	3.2(0.55)	5.75 <sup>***</sup>	49	<0.001
	Atm N <sub>5 year</sub>	-1.2(0.45)	-2.68 <sup>**</sup>		
<i>eH</i> <sub>1991-2012</sub>	Soil pH	3.2(0.40)	8.01 <sup>***</sup>	49	<0.001
	Atm N <sub>5 year</sub>	-1.2(0.32)	-3.69 <sup>***</sup>		
Legumes <sub>1903-2012</sub>	Soil pH	0.74(0.140)	5.34 <sup>***</sup>	67	<0.001
	Atm N <sub>5 year</sub>	-0.48(0.089)	-5.37 <sup>***</sup>		
Grass <sub>1903-2012</sub>	Soil pH	-0.56(0.078)	-7.04 <sup>***</sup>	67	<0.001
	Atm N <sub>5 year</sub>	0.14(0.041)	3.30 <sup>**</sup>		
Other <sub>1903-2012</sub>	Soil pH	0.40(-0.069)	5.83 <sup>***</sup>	69	<0.001

A normal distribution with an identity link was used for species number and *eH* and a binomial distribution with a logit link for the proportion of legume, grass and other (non-leguminous forbs), each analysed separately (over-dispersion was corrected for by estimating the dispersion parameter and using the *F* statistic to test for significance). For plant functional groups, data were available from 1903 but, because of changes in the sampling protocol, only data from the recent sampling period (1991–2012) were analysed for species richness and *eH*.

**Extended Data Table 4 | Effect of declining N deposition (measured either as a 3- or 5-year moving average), pH and N or P fertilizers on metrics of plant diversity**

Response variable	Explanatory variable	Estimate (s.e.)	F statistic	Degrees of freedom	P value
Species number	pH	0.20 (0.035)	44.2	366	<b>&lt;0.001</b>
	+N:48kg $\text{ha}^{-1}$ *	-0.92 (0.346)	11.5	27	<b>&lt;0.001</b>
	96kg $\text{ha}^{-1}$	-0.47 (0.210)			
	144kg $\text{ha}^{-1}$	-1.28 (0.467)			
	+Phosphorous	ns	2.6	26	0.120
$eH^i$	N Deposition <sub>3 year</sub>	-0.09 (0.028)	9.8	11	<b>0.010</b>
	pH	0.18 (0.044)	28.3	134	<b>&lt;0.001</b>
	+N:48kg $\text{ha}^{-1}$	-0.71 (0.291)	11.3	23	<b>&lt;0.001</b>
	96kg $\text{ha}^{-1}$	-0.47 (0.174)			
	144kg $\text{ha}^{-1}$	-1.14 (0.389)			
Proportion of Legumes	+Phosphorous	ns	0.4	21	0.514
	N Deposition <sub>5 year</sub>	-0.08 (0.033)	6.9	11	<b>0.023</b>
	pH	0.55 (0.151)	13.9	298	<b>&lt;0.001</b>
	+N:48kg $\text{ha}^{-1}$	-2.87 (1.739)	6.8	43	<b>&lt;0.001</b>
	96kg $\text{ha}^{-1}$	-2.33 (0.625)			
Proportion of Grass	144kg $\text{ha}^{-1}$	-5.08 (2.304)			
	+Phosphorous	2.23 (0.672)	11.1	23	<b>0.003</b>
	N Deposition <sub>5 year</sub>	-0.55 (0.185)	8.8	10	<b>0.014</b>
	pH	-0.85 (0.088)	110.4	115	<b>&lt;0.001</b>
	+N:48kg $\text{ha}^{-1}$	1.07 (0.265)	31.7	24	<b>&lt;0.001</b>
	96kg $\text{ha}^{-1}$	1.63 (0.489)			
	144kg $\text{ha}^{-1}$	3.04 (0.265)			
	+Phosphorous	ns	0.0	19	0.997
	N Deposition <sub>5 year</sub>	0.27 (0.122)	4.8	11	<b>0.050</b>

\*Effect size of additional N fertiliser expressed in relation to plots receiving no added nitrogen.

GLMMs were fitted to the data from all sub-plots sampled in the modern day period (1991–2012), with the exception of the one plot that receives organic manures, with plot and year input as random factors. A Poisson distribution with a log link was used for species richness and  $eH^i$  and a binomial distribution with a logit link for the proportion of legumes and grasses, each analysed separately. Where necessary, over-dispersion was corrected for by estimating the dispersion parameter and using the  $F$  statistic to calculate  $P$ . Two additional variables,  $\pm$ potassium and whether the data were from a transition plot, did not explain any variance in the response variables in any of the models.

**Extended Data Table 5 | GLMs fitted to relative biomass data from the Park Grass plot 9 (N<sub>2</sub>PKNMg)**

Plot 9/1					
Response variable	Explanatory variable	Estimate (s.e.)	<i>t</i> statistic	Degrees of freedom	<i>F</i> probability
Species number	Soil pH	3.5(0.74)	4.74 <sup>***</sup>	36	<0.001
	Atm N <sub>5 year</sub>	-3.2(0.60)	-5.31 <sup>***</sup>		
<i>eH</i>	Soil pH	2.6(0.47)	5.56 <sup>***</sup>	37	<0.001
Legumes	Atm N <sub>5 year</sub>	-0.94(0.130)	-7.19 <sup>***</sup>	37	<0.001
Grass	Soil pH	-0.66(0.165)	-3.99 <sup>***</sup>	36	<0.001
	Atm N <sub>5 year</sub>	0.58(0.109)	5.30 <sup>***</sup>		
Other	Soil pH	0.54(0.145)	3.69 <sup>***</sup>	37	<0.001
Plot 9/2					
Response variable	Explanatory variable	Estimate (s.e.)	<i>t</i> statistic	Degrees of freedom	<i>F</i> probability
Species number	Soil pH	1.7(0.48)	3.63 <sup>***</sup>	36	<0.001
	Atm N <sub>5 year</sub>	-2.9(0.38)	-7.69 <sup>***</sup>		
<i>eH</i>	Soil pH	1.3(0.28)	4.77 <sup>***</sup>	36	<0.001
	Atm N <sub>5 year</sub>	-0.8(0.22)	-3.62 <sup>***</sup>		
Legumes	Atm N <sub>5 year</sub>	-0.73(0.174)	-4.16 <sup>***</sup>	37	<0.001
Grass	Soil pH	-0.37(0.124)	-3.02 <sup>***</sup>	36	<0.001
	Atm N <sub>5 year</sub>	0.30(0.094)	3.20 <sup>***</sup>		
Other	Soil pH	0.74(0.121)	6.12 <sup>***</sup>	36	<0.001
	Rain <sub>t-1</sub>	0.002(0.0008)	2.22 <sup>*</sup>		

On plot 9/1, N was last applied as ammonium sulfate in 1989. Sub-plot d with very low pH and species diversity was excluded from the analysis. All data are for the recent sampling period, 1991–2012. A normal distribution with an identity link was used for species number and *eH* and a binomial distribution with a logit link for the proportion of legume, grass and other (non-leguminous forbs), each analysed separately (over-dispersion was corrected for by estimating the dispersion parameter and using the *F* statistic to test for significance).



**Extended Data Table 6 | GLMs fitted to relative biomass data from the Park Grass plot 14, (N\*<sub>2</sub>PKNaMg)**

Plot 14/1					
Response variable	Explanatory variable	Estimate (s.e.)	<i>t</i> statistic	Degrees of freedom	<i>F</i> probability
Species number	Atm N <sub>5 year</sub>	-2.0(0.29)	-6.70 <sup>***</sup>	50	<0.001
<i>eH</i>	Soil pH	-1.0(0.45)	-2.21 <sup>*</sup>	49	<0.001
	Atm N <sub>5 year</sub>	-0.8(0.24)	-3.41 <sup>***</sup>		
Legumes	Atm N <sub>5 year</sub>	-0.47(0.098)	-4.84 <sup>***</sup>	50	<0.001
Grass	Atm N <sub>5 year</sub>	0.26(0.075)	3.46 <sup>***</sup>	50	<0.001
Other	Rain <sub>t-1</sub>	0.003(0.0005)	5.44 <sup>***</sup>	50	<0.001
Plot 14/2					
Response variable	Explanatory variable	Estimate (s.e.)	<i>t</i> statistic	Degrees of freedom	<i>F</i> probability
Species number	Atm N <sub>5 year</sub>	-2.1(0.40)	-5.41 <sup>***</sup>	50	<0.001
<i>eH</i>	Soil pH	-1.5(0.57)	-2.61 <sup>*</sup>	49	<0.001
	Atm N <sub>5 year</sub>	-1.3(0.23)	-5.51 <sup>***</sup>		
Legumes	Atm N <sub>3 year</sub>	-1.50(0.225)	-6.65 <sup>***</sup>	50	<0.001
Grass	Atm N <sub>5 year</sub>	0.31(0.073)	4.20 <sup>***</sup>	50	<0.001
Other	Atm N <sub>5 year</sub>	-0.35(0.076)	-4.59 <sup>***</sup>	49	<0.001
	Rain <sub>t-1</sub>	0.003(0.0007)	4.08 <sup>***</sup>		

On plot 14/1, N was last applied as sodium nitrate in 1989. All data are for the recent sampling period, 1991–2012. A normal distribution with an identity link was used for species number and *eH* and a binomial distribution with a logit link for the proportion of legume, grass and other (non-leguminous forbs), each analysed separately (over-dispersion was corrected for by estimating the dispersion parameter and using the *F* statistic to test for significance).

Extended Data Table 7 | Significant responses of individual species to year at the level of the sub-plot

% Variance explained	3a	3b	3c	3d	9/1a	9/2a	9/1b	9/2b	9/1c	9/2c	9/1d	9/2d	14/1a	14/2a	14/1b	14/2b	14/1c	14/2c	14/1d	14/2d
P-value	20.6	20.5	24.8	25.3	28.8	40.9	40.9	46.0	52.4	29.1	44.4	8	23.1	33.1	20.6	25.9	31.7	26.5	36.9	23.6
	0.006	0.002	0.003	0.003	0.004	0.001	0.001	0.001	0.001	0.003	0.001	ns	0.004	0.003	0.005	0.003	0.001	0.001	0.001	0.004
<b>Monocots</b>																				
<i>Agrostis capillaris</i>	-*										+	+					+		+	
<i>Alopecurus pratensis</i>									+								+			
<i>Anthoxanthum odoratum</i>																				
<i>Arrhenatherum elatius</i>									+	+										-*
<i>Briza media</i>		+	+	+																
<i>Bromus hordeaceus</i>					+	+	+	+	+	+									+	
<i>Carex caryophyllaea</i>	+	+	+	+																
<i>Carex flacca</i>	+		+																	
<i>Dactylis glomerata</i>				-*					+	+										-*
<i>Festuca rubra</i>										+							+			
<i>Helictotrichon pubescens</i>					+				+											
<i>Holcus lanatus</i>												-*	+	+		+		+		
<i>Lolium perenne</i>					+			+					+	+		+		+	+	+
<i>Luzula campestris</i>			+	+																
<i>Poa pratensis</i>					+	+	+	+	+	+									+	
<i>Poa trivialis</i>		+							+	+										
<b>Forbs</b>																				
<i>Achillea millefolium</i>				+	+				+						+				+	
<i>Ajuga reptans</i>			+	+																
<i>Anthriscus sylvestris</i>						+	+			+										
<i>Cerastium fontanum</i>								+	+	+			+					+		+
<i>Conopodium majus</i>	-*								+	+										
<i>Heracleum sphondylium</i>													+	+					-*	
<i>Leontodon autumnalis</i>			+																	
<i>Leontodon hispidus</i>				+	+	+										+				
<i>Pilosella officinarum</i>			+	+																
<i>Pimpinella saxifraga</i>	+	+	+																+	
<i>Potentilla reptans</i>				+																
<i>Plantago lanceolata</i>				+	+	+	+	+	+	+				+	+	+	+	+		+
<i>Ranunculus acris</i>					+	+	+	+	+	+										
<i>Rumex acetosa</i>															+				+	+
<i>Taraxacum officinale</i> agg.																			+	+
<i>Tragopogon pratensis</i>			+					+	+	+							-*		-*	
<b>Legumes</b>																				
<i>Lathyrus pratensis</i>		+			+	+	+	+	+						+					
<i>Lotus corniculatus</i>	+	+	+																	
<i>Trifolium pratense</i>		+	+	+	+	+	+	+	+					+	+	+	+	+	+	+
<i>Trifolium repens</i>	+								+				+	+			+	+	+	+

GLMs were used to quantify species responses and precipitation before the first cut included as a covariate for all plots.

# Sex-dependent dominance at a single locus maintains variation in age at maturity in salmon

Nicola J. Barson<sup>1\*</sup>, Tutku Aykanat<sup>2\*</sup>, Kjetil Hindar<sup>3</sup>, Matthew Baranski<sup>4</sup>, Geir H. Bolstad<sup>3</sup>, Peder Fiske<sup>3</sup>, Céleste Jacq<sup>4</sup>, Arne J. Jensen<sup>3</sup>, Susan E. Johnston<sup>5</sup>, Sten Karlsson<sup>3</sup>, Matthew Kent<sup>1</sup>, Thomas Moen<sup>6</sup>, Eero Niemelä<sup>7</sup>, Torfinn Nome<sup>1</sup>, Tor F. Næsje<sup>3</sup>, Panu Orell<sup>7</sup>, Atso Romakkaniemi<sup>7</sup>, Harald Sægvog<sup>8</sup>, Kurt Urdal<sup>8</sup>, Jaakko Erkinaro<sup>7</sup>, Sigbjørn Lien<sup>1</sup> & Craig R. Primmer<sup>2</sup>

Males and females share many traits that have a common genetic basis; however, selection on these traits often differs between the sexes, leading to sexual conflict<sup>1,2</sup>. Under such sexual antagonism, theory predicts the evolution of genetic architectures that resolve this sexual conflict<sup>2–5</sup>. Yet, despite intense theoretical and empirical interest, the specific loci underlying sexually antagonistic phenotypes have rarely been identified, limiting our understanding of how sexual conflict impacts genome evolution<sup>3,6</sup> and the maintenance of genetic diversity<sup>6,7</sup>. Here we identify a large effect locus controlling age at maturity in Atlantic salmon (*Salmo salar*), an important fitness trait in which selection favours earlier maturation in males than females<sup>8</sup>, and show it is a clear example of sex-dependent dominance that reduces intralocus sexual conflict and maintains adaptive variation in wild populations. Using high-density single nucleotide polymorphism data across 57 wild populations and whole genome re-sequencing, we find that the vestigial-like family member 3 gene (*VGLL3*) exhibits sex-dependent dominance in salmon, promoting earlier and later maturation in males and females, respectively. *VGLL3*, an adiposity regulator associated with size and age at maturity in humans, explained 39% of phenotypic variation, an unexpectedly large proportion for what is usually considered a highly polygenic trait. Such large effects are predicted under balancing selection from either sexually antagonistic or spatially varying selection<sup>9,10</sup>. Our results provide the first empirical example of dominance reversal allowing greater optimization of phenotypes within each sex, contributing to the resolution of sexual conflict in a major and widespread evolutionary trade-off between age and size at maturity. They also provide key empirical evidence for how variation in reproductive strategies can be maintained over large geographical scales. We anticipate these findings will have a substantial impact on population management in a range of harvested species where trends towards earlier maturation have been observed.

The importance of balancing selection in maintaining variation in fitness-related traits, which are expected to be under strong selection, is a long-standing question in evolutionary biology<sup>7,11,12</sup>, with recent models suggesting that balancing selection may be particularly important in maintaining genetic variation<sup>9,13</sup>. Sexually antagonistic selection on traits with a shared genetic architecture where each sex is displaced from their optimal phenotype is one mechanism generating balancing selection<sup>2,3,6,9,14</sup>. Theoretical models predict that dominance reversals, where the dominant allele in one sex is recessive in the other, would greatly reduce constraints on the resolution of sexual conflict and may be particularly efficient at maintaining variation through heterozygote superiority across the sexes<sup>6,12,15</sup>, although this architecture has never been observed in the wild. A paucity of empirical examples with

known genetic architecture means that the evolutionary consequences of sexual conflict, particularly its importance in maintaining adaptive variation<sup>3,6,16</sup>, remains largely unknown<sup>14,16</sup>.

The age at which an individual reproduces is a critical point in its life history. Age at maturity affects fitness traits including survival, size at maturity and lifetime reproductive success<sup>17</sup>. Age at maturity in Atlantic salmon represents a classic evolutionary trade-off: larger, later-maturing individuals have higher reproductive success on spawning grounds<sup>18</sup>, yet also have a higher risk of dying before first reproduction<sup>17</sup>. Atlantic salmon reproduce in freshwater, with offspring migrating to sea to feed before returning to their natal river to spawn. The number of years spent at sea before spawning, namely their age at maturity, or 'sea age', has a dramatic impact on size at maturity, typically 1–3 kg and 50–65 cm after 1 year compared with 10–20 kg and >100 cm after 3 or more years<sup>19</sup>. Males mature earlier and at smaller size on average, whereas females mature later, with a stronger correlation between body size and reproductive success compared with males<sup>18</sup>. There is evidence for sex-specific selection patterns on age at maturity, as life-history strategies differ considerably between males and females<sup>18</sup>.

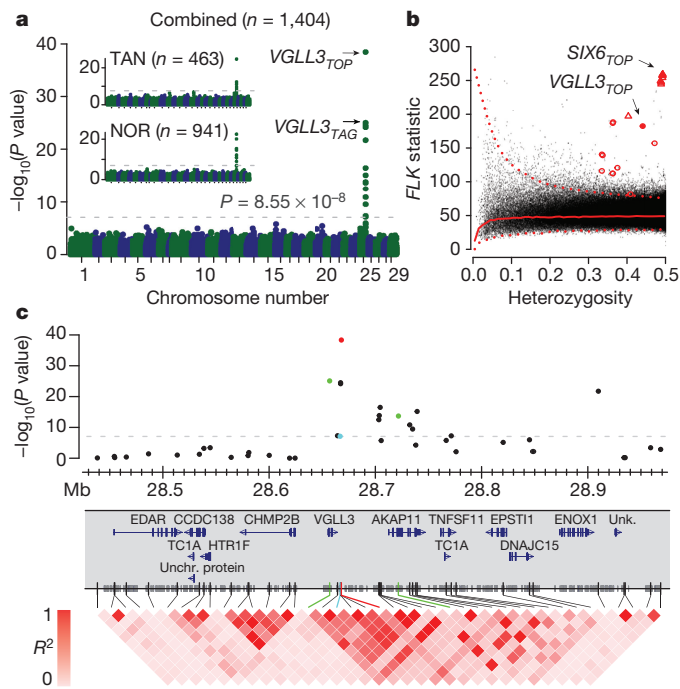
We investigated the genetic basis of age at maturity in Atlantic salmon using two independent data sets. The first, Tana (TAN), included two subpopulations from a large river system (Tana/Teno River; 68–70° N;  $n = 463$ ); the second, Norway (NOR), comprised 54 populations spanning the Norwegian coast from 59° N to 71° N, containing both Atlantic and Barents/White Sea phylogeographic lineages ( $n = 941$ ; NOR mean  $n$  per population = 17.4). Both data sets sampled geographically proximate populations with contrasting ages at maturity (Extended Data Fig. 1, Supplementary Information and Extended Data Table 1). Genome-wide association studies (GWAS) for age at maturity were conducted within both data sets using 208,704 single nucleotide polymorphisms (SNPs) (Supplementary Note). A region spanning approximately 100 kb on chromosome 25 was strongly associated with age at maturity in both data sets (GWAS;  $P < 1 \times 10^{-20}$ ; Fig. 1a, c and Extended Data Fig. 2) explaining 39.4% (standard error 1.1%) of the total phenotypic variation. This association was further validated in a phylogeographically distant Baltic Sea data set (BAL) ( $P < 9.74 \times 10^{-8}$ ; Extended Data Fig. 3), confirming that the region is evolutionarily conserved across all European lineages. The region included two candidate loci (Fig. 1c and Extended Data Fig. 4a), *VGLL3* and A-kinase anchor protein 11 (*AKAP11*). *VGLL3* is a transcription cofactor with a role in adipogenesis as a negative regulator of terminal adipocyte differentiation, and its expression is correlated with body weight and gonadal adipose content in mice<sup>20</sup>. *VGLL3* has also been associated with age at menarche<sup>21</sup> and pubertal height growth in humans<sup>22</sup>, indicating a remarkably high level of

<sup>1</sup>Centre for Integrative Genetics (CIGENE), Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, NO-1432 Ås, Norway. <sup>2</sup>Department of Biology, University of Turku, FI-20014, Finland. <sup>3</sup>Norwegian Institute for Nature Research (NINA), NO-7485 Trondheim, Norway. <sup>4</sup>Nofima - Norwegian Institute of Food, Fisheries and Aquaculture Research, NO-1431 Ås, Norway. <sup>5</sup>Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK. <sup>6</sup>AquaGen, NO-7462 Trondheim, Norway. <sup>7</sup>Natural Resources Institute Finland, Oulu, FI-90014, Finland.

<sup>8</sup>Radgivende Biologer, NO-5003 Bergen, Norway.

\*These authors contributed equally to this work.

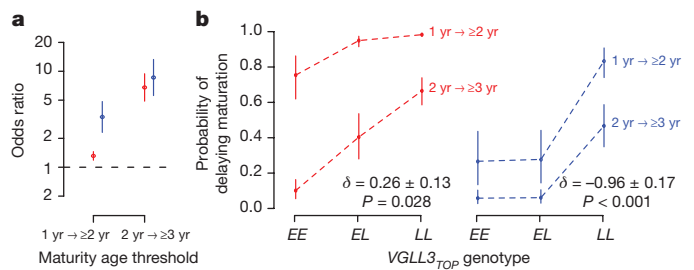




**Figure 1 | Genetic mapping of age at maturity and divergence across populations.** **a**, GWAS for age at maturity for the TAN and NOR data sets combined. Insets show the two data sets independently (Extended Data Fig. 2). **b**, Signatures of spatially divergent selection using the FLK  $F_{ST}$  outlier test (56 populations, total  $n = 1,404$ ). Solid and dashed lines indicate the smoothed median and 99.5% quantile of the neutral distribution, respectively. Ten SNPs flanking the *VGLL3*<sub>TOP</sub> and *SIX6*<sub>TOP</sub> SNPs (filled symbols) are marked with red circles and triangles, respectively. ( $P_{VGLL3TOP} = 1.44 \times 10^{-15}$  and  $P_{SIX6TOP} \approx 0$ ). **c** The gene model and linkage disequilibrium plot of the  $\sim 0.5$  Mb region around the significant region on chromosome 25. Notable SNPs are colour coded with red (*VGLL3*<sub>TOP</sub>), blue (*VGLL3*<sub>IHS</sub>) and green (SNPs tagging missense mutations in *VGLL3* and *AKAP11*). Shorter tick marks in the SNP axis indicate re-sequencing variants.

functional conservation. Age at menarche is associated with adiposity in humans<sup>21,22</sup>, and puberty in fish is linked to the absolute level or rate of accumulation of lipid reserves<sup>23</sup>. Threshold levels of fat reserves at critical times of year are thought to control the initiation of maturation in salmon<sup>24</sup>. Therefore, *VGLL3* may serve to regulate the interaction between fat reserves (adiposity) and maturation in salmon, in a similar manner to mammals, and is a strong candidate gene for age at maturity. *AKAP11* is expressed throughout spermatogenesis and is important for mature sperm motility<sup>25</sup>. Targeted whole genome re-sequencing of 32 individuals from seven populations revealed two missense mutations in *VGLL3* in strong linkage disequilibrium with a nearby highly associated genic SNP (*VGLL3*<sub>TAG</sub>) and with each other (Met54Thr–*VGLL3*<sub>TAG</sub>  $r^2 = 1$ ; Asn323Lys–*VGLL3*<sub>TAG</sub>  $r^2 = 0.72$ ; Met54Thr–Asn323Lys  $r^2 = 0.72$ ; Extended Data Fig. 4a) and confirmed a missense SNP had been genotyped in *AKAP11* (Fig. 1c and Extended Data Fig. 4a). A test for predicting changes in protein structure/function (PolyPhen2, see Methods) strongly supported two of these mutations having an effect on phenotype, owing to high evolutionary conservation of the codons (*VGLL3* Asn323Lys, naive Bayes posterior probability = 0.976, sensitivity = 0.76, specificity = 0.96; *AKAP11* Val214Met, naive Bayes posterior probability = 0.716, sensitivity = 0.86, specificity = 0.92).

A second genomic region, which spans 250 kb on chromosome 9, was strongly associated with age at maturity ( $P < 10^{-20}$ ; Extended Data Fig. 2a, Extended Data Fig. 4b–c and Supplementary information), but was no longer significant after population stratification correction (Extended Data Fig. 2). This signal is likely to represent between-population variation in a correlated trait, size at maturity

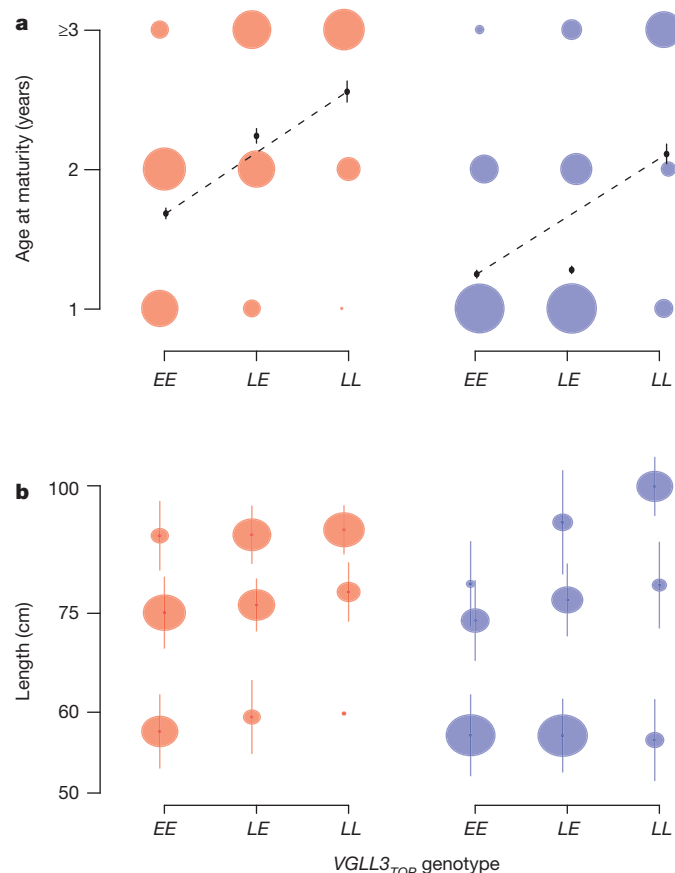


**Figure 2 | Genetic architecture of age at maturity in the *VGLL3*<sub>TOP</sub> locus.** **a**, Odds ratio (median) between the alternative homozygous genotypes for delaying maturation in females ( $n = 693$ , red) and males ( $n = 711$ , blue). Error bars are 50% sampling quantiles (100,000 parametric permutations). All odds are significantly different from 1 ( $P < 0.001$ ). **b**, Probability of delaying maturation as a function of *VGLL3*<sub>TOP</sub> genotype in females (red) and males (blue). Dominance estimates on the liability scale are given for each sex (see also Extended Data Fig. 5). Note that the 2–3 year category in females did not deviate significantly from additivity on the observed scale.

(Supplementary Information, Extended Data Fig. 2c–e). The core haplotype included a transcription factor of the hypothalamus–pituitary–gonadal axis, *SIX6*, associated with size and age at maturity in humans<sup>21</sup> and a conserved non-coding element that aligns to a candidate distal forebrain enhancer of *SIX6*<sup>26</sup> (Extended Data Fig. 4b–c and Supplementary information). Both genome regions also exhibited strong signals of spatially divergent selection across populations (FLK  $F_{ST}$  outlier test,  $P < 10^{-15}$ ; Fig. 1b).

Two alleles at the most highly associated SNP in the *VGLL3* locus (*VGLL3*<sub>TOP</sub>) conferred either early (E) or late (L) maturation. LL individuals had significantly higher odds ratios for delaying maturation, particularly for older maturity ages (Fig. 2a) and were predicted to mature, on average, 0.87 (females) and 0.86 (males) years later than EE individuals (Fig. 3a); a remarkable shift considering the average sea age at maturity in salmon is 1.6 years (population range averages 1.0–2.6)<sup>19</sup>. This locus also influenced size of individuals with the same age at maturity in both sexes, with a genotype-by-maturity interaction in males: for example, length = 100 and 80 cm, for LL and EE males maturing after 3 years at sea, respectively ( $P = 0.006$ ; Fig. 3b and Supplementary Table 1). In addition, there were striking differences in dominance patterns between the sexes: in females the L allele was partly dominant across threshold categories ( $\delta = 0.26 \pm 0.13$ ,  $P = 0.028$ ), whereas in males the E allele was completely dominant ( $\delta = -0.96 \pm 0.17$ ,  $P < 0.001$ ; Figs 2b and 3a, Extended Data Fig. 5 and Extended Data Table 2), providing a compelling mechanism contributing to the larger proportion of males exhibiting an early maturing phenotype compared with females<sup>18</sup>. Variation at *VGLL3* was maintained in all but 1 of the 54 NOR populations, with all populations characterized by large salmon maintaining intermediate allele frequencies, consistent with balancing selection<sup>7</sup> (Extended Data Table 1 and Extended Data Fig. 2c). Given that a large proportion of variation in age at maturity (and subsequent body size) is governed by a single locus of large effect, such sex-specific trade-offs with a shared genetic basis could effectively maintain genetic variation under varying patterns of dominance between the sexes<sup>6</sup>.

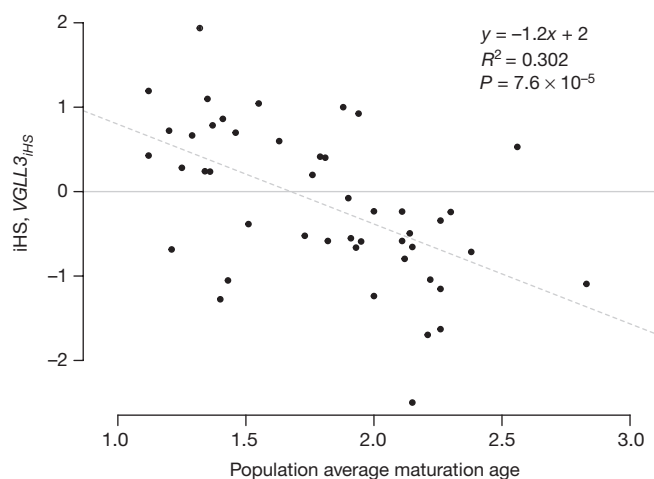
The large effect sizes of the *VGLL3* alleles are consistent with evolutionary theory, which predicts that beneficial alleles of intermediate-to-large effects are likely to be maintained under balancing selection, particularly when their phenotypic and fitness effects differ between the sexes<sup>9</sup>. Evolution towards complex traits controlled by fewer loci with larger effects is also predicted where gene flow between environments with different trait optima results in balancing selection<sup>10</sup>. We investigated whether spatially varying selection on *VGLL3*, suggested by the FLK  $F_{ST}$  analysis (Fig. 1b), was consistent with selection towards local optima. We found a strong effect of a population's average age at maturity on



**Figure 3 | Effect of the *VGLL3*<sub>TOP</sub> genotype on age at maturity and size.** **a**, Age at maturity (in years) of females ( $n = 693$ , red) and males ( $n = 711$ , blue) in relation to *VGLL3*<sub>TOP</sub> genotype. Circle areas are proportional to sample size. Black dots indicate predicted average sea age using logit transformation model, and error bars are 50% sampling quantiles (10,000 parametric permutations). **b**, *VGLL3*<sub>TOP</sub> genotypic effect on size within maturation age classes. The average length (in centimetres) of females (red) and males (blue) maturing after 1, 2 or 3 years are indicated by the lower, middle and upper three dots, respectively. Length (in centimetres) on the y axis is log scaled and corrected for population effects. Circle diameters are proportional to sample size, and lines indicate sample s.d.

the integrated haplotype score (iHS), a measure of the amount of extended haplotype homozygosity around one allele of an SNP relative to the alternative allele (slope =  $-1.18 \pm 0.27$  standard error per year,  $R^2 = 0.302$ ,  $P = 7.6 \times 10^{-5}$ ; Fig. 4, Extended Data Fig. 6 and Supplementary Information). In populations with an older average age at maturity, there were relatively higher levels of extended homozygosity around the *L* allele compared with the *E* allele, while the pattern was the opposite in populations with younger average age at maturity (Fig. 4 and Extended Data Fig. 6). This result suggests a systematic shift in selection pressure for earlier/late maturation alleles coincident with the population's average age at maturity and is consistent with divergent selection among populations towards local optima (Supplementary Information), and an effect of gene flow on the observed genetic architecture<sup>10</sup>.

Our results reveal a major effect locus determining age at maturity in Atlantic salmon. The large effect of this locus is remarkable given that age at maturity is generally considered a classic polygenic trait<sup>21</sup>. A shared gene controlling age at maturity between mammals and a teleost fish provides evidence for evolutionary conservation across large taxonomic distances for a life-history trait, as observed for morphological characters<sup>27</sup>. Our results provide the first empirical example of dominance reversal allowing greater optimization of phenotypes within each sex. Partial dominance of the higher fitness allele in each sex can result in a net effect of heterozygote superiority



**Figure 4 | Relationship between population iHS score (46 populations, 32 haplotypes per population) and average maturation age of each population for the *VGLL3*<sub>IHS</sub> locus.** iHS = 0 (no haplotype length difference) is marked with a horizontal grey line. Positive iHS values indicate longer haplotype blocks, and therefore stronger selection around the *E* allele in a population relative to the *L* allele and vice versa for negative iHS values.

across the sexes, and thus maintain stable polymorphisms<sup>6</sup>. In common with many other species, Atlantic salmon lack heteromorphic sex chromosomes<sup>28</sup>, which precludes the use of the X chromosome to protect sexual conflict polymorphisms. Sex-dependent dominance removes the restrictive conditions on maintaining conflict alleles on autosomes, making sexually antagonistic polymorphism more likely to be maintained on autosomes than on the X chromosome<sup>3,6</sup>. In line with our results, restrictive conditions on the maintenance of variation by balancing selection suggest fewer, large effect loci will control traits under both sexual antagonism and spatially varying selection<sup>7,9,10</sup>. The discovery of a major locus affecting age at maturity will have a substantial impact on population management of Atlantic salmon, where a decrease in the frequency of late maturation has been observed in many populations<sup>29</sup>, and potentially other exploited species showing comparable shifts towards earlier maturation<sup>30</sup> if this architecture is similar in other species.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 12 August; accepted 7 October 2015.**

**Published online 4 November 2015.**

- Bonduriansky, R. & Chenoweth, S. F. Intralocus sexual conflict. *Trends Ecol. Evol.* **24**, 280–288 (2009).
- Lande, R. Sexual dimorphism, sexual selection, and adaptation in polygenic characters. *Evolution* **34**, 292–305 (1980).
- Rice, W. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38**, 1416–1424 (1984).
- Fisher, R. A. *The Genetical Theory of Natural Selection* 139–142 (Oxford Univ. Press, 1930).
- van Doorn, G. S. Intralocus sexual conflict. *Ann. NY Acad. Sci.* **1168**, 52–71 (2009).
- Fry, J. D. The genomic location of sexually antagonistic variation: some cautionary comments. *Evolution* **64**, 1510–1516 (2010).
- Turelli, M. & Barton, N. H. Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and  $G \times E$  interactions. *Genetics* **166**, 1053–1079 (2004).
- Schaffer, W. M. in *Evolution Illuminated: Salmon and Their Relatives* (eds Stearns, S. & Hendry, A.) 20–51 (Oxford Univ. Press, 2004).
- Connallon, T. & Clark, A. G. Balancing selection in species with separate sexes: insights from Fisher's geometric model. *Genetics* **197**, 991–1006 (2014).
- Savolainen, O., Lascoux, M. & Merilä, J. Ecological genomics of local adaptation. *Nature Rev. Genet.* **14**, 807–820 (2013).
- Dobzhansky, T. A review of some fundamental concepts and problems of population genetics. *Cold Spring Harb. Symp. Quant. Biol.* **20**, 1–15 (1955).
- Mokkonen, M. et al. Negative frequency-dependent selection of sexually antagonistic alleles in *Myodes glareolus*. *Science* **334**, 972–974 (2011).

13. Sellis, D., Callahan, B. J., Petrov, D. A. & Messer, P. W. Heterozygote advantage as a natural consequence of adaptation in diploids. *Proc. Natl Acad. Sci. USA* **108**, 20666–20671 (2011).
14. Connallon, T. & Clark, A. G. Evolutionary inevitability of sexual antagonism. *Proc. R. Soc. Lond. B* **281**, 20132123 (2014).
15. Kidwell, J. F., Clegg, M. T., Stewart, F. M. & Prout, T. Regions of stable equilibria for models of differential selection in the two sexes under random mating. *Genetics* **85**, 171–183 (1977).
16. Pennell, T. M. & Morrow, E. H. Two sexes, one genome: the evolutionary dynamics of intralocus sexual conflict. *Ecol. Evol.* **3**, 1819–1834 (2013).
17. Stearns, S. C. Life history evolution: successes, limitations, and prospects. *Naturwissenschaften* **87**, 476–486 (2000).
18. Fleming, I. A. & Einum, S. in *Atlantic Salmon Ecology* 33–65 (Wiley-Blackwell, 2011).
19. Hutchings, J. A. & Jones, M. E. B. Life history variation and growth rate thresholds for maturity in Atlantic salmon, *Salmo salar*. *Can. J. Fish. Aquat. Sci.* **55** (Suppl. 1), 22–47 (1998).
20. Halperin, D. S., Pan, C., Lusi, A. J. & Tontonoz, P. Vestigial-like 3 is an inhibitor of adipocyte differentiation. *J. Lipid Res.* **54**, 473–481 (2013).
21. Perry, J. R. B. et al.; Australian Ovarian Cancer Study; GENICA Network; kConFab; LifeLines Cohort Study; InterAct Consortium; Early Growth Genetics (EGG) Consortium. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
22. Cousminer, D. L. et al.; ReproGen Consortium; Early Growth Genetics (EGG) Consortium. Genome-wide association and longitudinal analyses reveal genetic loci linking pubertal height growth, pubertal timing and childhood adiposity. *Hum. Mol. Genet.* **22**, 2735–2747 (2013).
23. Taranger, G. L. et al. Control of puberty in farmed fish. *Gen. Comp. Endocrinol.* **165**, 483–515 (2010).
24. Thorpe, J., Mangel, M., Metcalfe, N. & Huntingford, F. Modelling the proximate basis of salmonid life-history variation, with application to Atlantic salmon, *Salmo salar* L. *Evol. Ecol.* **12**, 581–599 (1998).
25. Reinton, N. et al. Localization of a novel human A-kinase-anchoring protein, hAKAP220, during spermatogenesis. *Dev. Biol.* **223**, 194–204 (2000).
26. Lee, B. et al. Direct transcriptional regulation of Six6 is controlled by SoxB1 binding to a remote forebrain enhancer. *Dev. Biol.* **366**, 393–403 (2012).
27. Flatt, T. & Heyland, A. (eds) *Mechanisms of Life History Evolution: The Genetics and Physiology of Life History Traits and Trade-Offs* (Oxford Univ. Press, 2011).
28. Woram, R. A. et al. Comparative genome analysis of the primary sex-determining locus in salmonid fishes. *Genome Res.* **13**, 272–280 (2003).
29. Chaput, G. Overview of the status of Atlantic salmon (*Salmo salar*) in the North Atlantic and trends in marine mortality. *ICES J. Mar. Sci.* **69**, 1538–1548 (2012).
30. Allendorf, F. W. & Hard, J. J. Human-induced evolution caused by unnatural selection through harvest of wild animals. *Proc. Natl Acad. Sci. USA* **106** (Suppl. 1), 9987–9994 (2009).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank L. Andersson, T. F. Hansen and H. Granroth-Wilding for commenting on earlier drafts of the manuscript. We also acknowledge the numerous fishers who contributed scales and phenotypic information. We thank J. Haantie, J. G. Jensås, J. Kuusela, I. Torvi and G. Østborg for scale measurements, T. Andersstuen, T. Balstad, L. Birkeland Eriksen, S. Karoliussen, J. Kuismin, M. Lindqvist, T. Pajula, K. Salminen, K. Söstar, M. Spets and K. Vagonyte-Hallan for laboratory assistance, M. Ellmen, O. Guttorm, T. Kanninen, A. Koskinen, T. Pöyhönen and S. Uusi-Heikkilä for sampling assistance, and T. Mulugeta for informatics support. Bioinformatic analyses used resources at the Finnish Centre for Scientific Computing, the Abel Cluster, owned by the University of Oslo and the Norwegian Metacenter for High Performance Computing, and operated by the Department for Research Computing at the University of Oslo IT Department and the Orion Computing Cluster at CIGENE. This study was funded by the Finnish Academy (grants 137710, 141231, 272836, 284941), the Research Council of Norway (QuantEscape, grant 216105 and RCN-project 221734/O30) and by AquaGen (SNP array development).

**Author Contributions** C.R.P., S.L., N.J.B., T.A. and K.H. conceived the study. C.R.P., S.L., N.J.B., T.A., K.H., C.J., S.K. and S.E.J. designed the experiments. T.M. led the development of the 220K SNP array, and M.K. and T.N. generated and conducted bioinformatics on the molecular data. K.H., P.F., A.J.J., T.F.N., H.S., K.U., J.E., P.O., A.R. and E.N. coordinated the collection of phenotypic data. T.A., N.J.B., M.B., G.H.B., S.K. and C.J. analysed the data. N.J.B., T.A. and C.R.P. wrote the manuscript. All authors read and commented on the manuscript. C.R.P. and S.L. contributed equally as senior authors.

**Additional Information** Details of the SNPs used in the study have been deposited in dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>) under accession numbers ss1867919552–ss1868858426, and re-sequencing data have been deposited in EMBL Nucleotide Sequence Database (European Nucleotide Archive) under accession number PRJEB10744. SNP genotype and phenotype data and detailed DNA sequence information of the main candidate gene regions are available in Dryad (<http://dx.doi.org/10.5061/dryad.23h4q>). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.R.P. (Craig.primmer@utu.fi) or S.L. (sigbjorn.lien@nmbu.no).



## METHODS

**Study design and study material.** Norway data set (NOR). Individuals were sampled from populations spanning the Norwegian coast from the Skagerrak in the south (59° N) to the Barents Sea in the north (71° N). In total 54 populations were sampled ( $n = 941$ ), including 11 populations within the Barents/White Sea phylogeographic group<sup>31</sup>, the remainder belonging to the Atlantic phylogeographic group (Extended Data Table 1). Samples were initially filtered to remove any individuals with possible aquaculture escapee ancestry<sup>32</sup>.

Tana data set (TAN). Two sub-populations, occurring in sympatry<sup>33,34</sup> in the mainstem Tana River, were subjected to in-depth within-population sampling ( $n = 463$ ). Scales were collected from salmon harvested by local fishermen using a variety of methods (nets, rod) from 2001 to 2003. In total, 326 and 137 individuals from each sub-population were included.

Baltic Sea data set. The BAL data set included 114 individuals from the Tornio river (66–69° N, 19–25° E) (Extended Data Fig. 1, Extended Data Table 1). This population belongs to the phylogeographically distinct Baltic Sea lineage<sup>31</sup>. Scales were collected from individuals harvested by trained anglers from 2003 to 2005. Storage and phenotypic measurements were as described for the TAN samples.

In total, the study included 1,518 Atlantic salmon individuals after initial data filtering that removed low-quality samples and individuals with signs of aquaculture escapee ancestry (Extended Data Table 1).

**Phenotypic measurements.** Length at capture (LEN) and weight at capture (WGT) were recorded during sampling. The sex (SEX) of most individuals was determined genetically in the NOR and TAN data sets<sup>35</sup>, while phenotypic sex determination was used for a small subset of samples (15% and 0.3% for TAN and NOR data sets respectively). Similar to tree rings, scale growth in fishes is commonly used to infer individual growth and age<sup>36,37</sup>. Growth, freshwater age (that is, age before sea migration, FW Age), and years spent at sea before first sexual maturation and spawning, referred to here as age at maturity (Mat Age), were inferred from scales using internationally agreed guidelines for Atlantic salmon scale reading<sup>38</sup>. Early life-history growth traits and size were assessed for their influence on age at maturity as it has been shown that freshwater growth may be negatively correlated with seawater growth<sup>39</sup> and that freshwater size may be positively correlated with age at maturity<sup>40</sup>. Freshwater size (FWS), freshwater growth (FWG), as well as first year growth at sea (SWG) were derived from the scale data, and used as independent variables throughout the analyses. Freshwater size is the log radius of the scale from the scale centre to the end of the freshwater growth period, and growth at sea is the log radius of the scale from the end of the freshwater growth to the end of the first winter annulus at sea. Freshwater growth is dependent on freshwater size and negatively on freshwater age. The residuals of the linear regression between freshwater size and freshwater age were further corrected for freshwater age to obtain the freshwater growth metric. As expected, a model where freshwater size was nested within freshwater age explained 97% of freshwater growth (analysis of variance,  $P < 10^{-16}$ ). Size at the end of first year at sea (SWS) was completely dependent on freshwater size and growth at sea, and therefore was not explored as an independent variable to avoid co-linearity.

**Genotyping and data filtering.** SNP array details. A custom 220,000 SNP Affymetrix Axiom array was used to genotype samples according to the manufacturer's instructions with a GeneTitan genotyping platform (Affymetrix). The SNPs on this array were a subset of those included on the 930K XHD Ssal array (dbSNP accession numbers ss1867919552–ss1868858426), and were chosen for maximum informativeness on the basis of their SNPfilter performance (SNPfilter, version 1.4, Affymetrix), minor allele frequency (maf) in aquaculture samples ( $\text{maf} > 0.05$ ) and physical distribution. The ascertainment bias of this array for wild Norwegian salmon is expected to be low because of the recent founding of the aquaculture population from a large number ( $n = 40$ ) of Norwegian salmon populations<sup>41</sup>. Within each population, the order of samples was randomized with respect to age at maturity, and the genotype calling was conducted by an automated pipeline without knowledge of age at maturity status of each individual. No statistical methods were used to predetermine sample size.

Raw genotyping data were analysed using the Linux-based APT pipeline applying best practice thresholds (contrasts quality control (DQC) threshold, 0.82; STEP1, 0.97). After the initial sample filtering, markers with low maf ( $< 0.01$ ) and/or call rate ( $< 0.97$ ) were filtered out using the *check.marker* function in the GenABEL package (version 1.8.8)<sup>42</sup> in the R environment (version 3.1.0)<sup>43</sup> for the GWAS. The same data parameters were used for the data set for phasing except that no maf threshold was set (see below). We did not perform a Hardy–Weinberg equilibrium test, since the data set contained individuals from multiple populations. After the filtering steps, 208,704 SNPs in 29 linkage groups remained in the analysis. An additional filtering was performed separately for the BAL data set ( $\text{maf} < 0.01$ ), resulting in 167,410 SNPs remaining in this data set for the GWA analysis.

**Model selection and GWAS of maturation age.** We performed a GWAS of age at maturity using an additive cumulative proportional odds model with the R package ordinal<sup>44</sup>, where the age at maturity propensity of a genotype was evaluated using a logit link model, and a flexible threshold structure (Supplementary Information). In addition to analysing the NOR and TAN data sets separately, model selection was performed with the two data sets combined (NOR + TAN,  $n = 1,404$ ) including geographical coordinates as parameters. Model selection was not performed for the BAL samples, for which we did not have freshwater phenotypic information available, and we performed GWA analysis without phenotypic co-variables (that is, BASIC model). Supplementary Table 2 lists the details of the model selection parameters for TAN, NOR and the combined (TAN + NOR) data sets.

We performed all GWA analyses with the BASIC model in addition to the FULL model to assess the effect of inclusion of covariates on the model (for example ref. 45; see Extended Data Fig. 2). The GWA analysis used a model comparison approach, where the effect of the SNP loci was evaluated by comparing the likelihood of the observational level model (as above) to the additive genetic model with SNP loci as covariates. Genome-wide statistical significance was adjusted for multiple comparisons and genomic inflation ( $\lambda$ ) for each analysis ( $P = 0.05 \times n_{\text{SNP}} \times \lambda$ ). Specific significance thresholds are listed in Extended Data Fig. 2.

To account for population stratification, we fitted the same model as above but also included principal components derived from the genomic kinship matrix as fixed factors. Principle components were added sequentially in the model until origin of population no longer explained a significant portion of genetic variance across SNPs (Supplementary Information). The optimal number of principal components was one for TAN, and 14 for NOR and the combined (TAN + NOR) data sets (Extended Data Fig. 2). Population structure in the BAL data set was corrected using two principal components as fixed factors, which reduced the  $\lambda$  value to 1.07 (Extended Data Fig. 3). We also compared association statistics of BAL ( $n = 114$ ) and the combined data (TAN + NOR,  $n = 1,404$ ) post hoc, to assess the magnitude of the effect of sample size on the association statistic of the *VGLL3<sub>TOP</sub>* locus. The TAN + NOR data set was re-sampled 100,000 times with an equivalent sample size and age at maturity structure to the BAL data set. The observed association statistic for the *VGLL3<sub>TOP</sub>* locus in the BAL set was similar to that in the TAN + NOR re-sampled data sets (Kolmogorov–Smirnov test,  $P = 0.51$ , Extended Data Fig. 3) indicating that the lower  $P$  value in BAL is probably caused by the lower sample size.

**Identifying signatures of spatially divergent selection:  $F_{ST}$  outlier test.** We used an extension of the Lewontin and Krakauer test, the  $FLK F_{ST}$  outlier test, which uses population trees (using Reynold's genetic distances and neighbour joining algorithm) to estimate expected neutral evolution (null) among populations<sup>46</sup>. This method has been shown to perform well under different demographic scenarios<sup>47</sup>. The empirical null distribution of SNPs was identified using the estimated population tree and 100,000 simulations.

**Mode of inheritance and effect sizes of age at maturity loci.** We detailed the genetic architecture of the loci associated with age at maturity by evaluating the likelihood of several inheritance models. In addition to simple additive and dominance models, we also tested various models where dominance inheritance was modelled conditioned on sex. Extended Data Table 2 lists the details of each model. Models were compared using an information-theoretic approach, where the model with the lowest Akaike information criterion was accepted as the optimal model explaining the data. The coefficient of the optimal model for the *VGLL3<sub>TOP</sub>* locus is given in Supplementary Table 3.

The patterns of dominance in the *VGLL3<sub>TOP</sub>* locus were investigated in detail, at the unobserved liability scale. Deviations from additivity were tested for each sex separately. In addition, deviation of genetic architectures between the sexes was also tested. For these tests, we used genotype coefficients ( $\beta_{\text{genotype}}$ ) and the standard errors obtained by the threshold model (Supplementary Table 3). We first standardized the coefficient to  $[0, 1]$  range, such that  $(\beta_{LL} + \beta_{EE})/2 = 0.5$  (that is, the average of the homozygote genotypes). Ten thousand parametric permutations were drawn from genotype coefficients, and the additive expectation (that is, null) was calculated as  $(\beta_{LL} + \beta_{EE})/2$ , which was compared with  $\beta_{EL}$  (heterozygote). For direct comparisons between sex dominance patterns, test statistics were reported as the proportion of samples deviating from the null ( $H_{\text{null}}: \beta_{EL\text{female}} = \beta_{EL\text{male}}$ ) in one direction.

**The proportion of variation explained by the *VGLL3<sub>TOP</sub>* locus.** To estimate the proportion of variance in age at maturity explained by the *VGLL3<sub>TOP</sub>* genotype, we employed an alternative modelling framework, where the response variable (Mat Age) is expressed on the logit scale with the  $y = \text{Mat age}/(1 + \text{Mat Age})$  transformation. Advantages of this transformation, where  $\text{logit}(\text{Mat Age}/(1 + \text{Mat Age}))$  is equal to  $\log(\text{Mat Age})$  and thus the coefficients are on the (log) observational scale, are that it (1) conveniently allows quantification of the examined variation in relation to total variation, and (2) enables quantification of effect

size with a straightforward interpretation. We employed this framework to the NOR, TAN and combined data sets with the same specification as in the FULL model (including principal components as fixed factors) using glmer function in the lme4 package (version 1.1-7)<sup>48</sup> in R, with a binomial link and bobyqa optimizer. The model provided fits comparable to the additive cumulative proportional odds model, where,  $R^2_{MF}$  and  $R^2_{CS}$  were 0.19 and 0.28 for the TAN, 0.13 and 0.17 for NOR, and 0.03 and 0.12 for the combined data sets, respectively (see Supplementary Information and Supplementary Table 4 for model details).

In addition to age at maturity, we also analysed the genotypic effect of *VGLL3<sub>TOP</sub>* locus on variation in size at return, using analysis of variance, and after accounting for maturation age. We modelled sexes separately because of the sex-dependent genetic architecture; the genotype-by-maturity age interaction term was included in the models. Population effects were calculated with a similar framework as in the GWA analysis such that the same number of principal components were used here to correct for the population inflation factor.

**Estimation of the coefficient of determination.** McFadden's pseudo- $R^2$  and Cox and Snell's pseudo- $R^2$ , which are two commonly used metrics that generalize the OLS  $R^2$  framework to logistic models, were used to evaluate the goodness of fit of the model. The likelihood differences between the alternative (that is, with *VGLL3<sub>TOP</sub>* genotypes) and the null hypothesis were McFadden's  $R^2$  ( $R^2_{MF}$ ) =  $1 - (\log L_{alt}/\log L_{null})$ , and Cox and Snell's  $R^2$  ( $R^2_{CS}$ ) =  $1 - (L_{null}/L_{alt})^{2/n}$ , where  $n$  equals the number of observations. We used the optimum model explaining genetic architecture as  $H_{alt}$  (Extended Data Table 2), and excluded only genotype terms from the model for the  $H_{null}$ . For TAN,  $R^2_{MF}$  and  $R^2_{CS}$  were 0.17 and 0.23. For NOR,  $R^2_{MF}$  and  $R^2_{CS}$  were 0.10 and 0.15 and the equivalent values for the combined data set were 0.12, 0.17, respectively.

**Genome re-sequencing and functional variant detection.** Thirty-two wild Atlantic salmon were selected for whole genome re-sequencing from seven populations (three from the Barents/White Sea and four from the Atlantic phylogeographical groups; see also Extended Data Table 1 and Supplementary Table 5). Three individuals per population were re-sequenced, except for the Tana sub-populations, where 14 individuals were re-sequenced. DNA was isolated from 14 adipose fin-clips (stored in ethanol) and 18 scale samples collected in 2012 and 2013 (stored in paper envelopes) using Qiagen DNAeasy kits according to the manufacturer's recommendations. DNA was quantified using Qubit fluorometry (Invitrogen).

For high-quality DNA derived from adipose tissue and two scale-derived DNA extractions that had high DNA quantity and quality, sequencing libraries were produced using a TruSeq DNA PCR-free Library Preparation Kit. Libraries for the remaining 16 scale-derived DNA extracts were prepared using a TruSeq Nano DNA Library Preparation Kit. The main motivation for this difference was to select kits most suited to available sample quantities: both kits use mechanical fragmentation (Covaris), thus limiting a bias caused by using a mixture of enzymatic and mechanical approaches. Library preparations were performed according to the manufacturer's instructions (Supplementary Table 5). All libraries were subjected to a fragment size selection (mode = 350 base pairs) and sequenced to generate  $2 \times 125$  nucleotide paired-end reads using an Illumina HiSeq 2500 platform. Sample preparation and sequencing were performed by the Norwegian Sequencing Centre, Ullevål (Oslo, Norway). Only reads passing Illumina's chastity filter were used in subsequent analysis. We further used FastQC to assess sequencing quality, passing lanes where the per-base quality score box plot indicated bases 1–110 having  $>Q20$  for  $>75\%$  of the reads. All lanes passed the quality criteria.

Reads were mapped to the salmon reference genome (National Center for Biotechnology Information Whole Genome Shotgun (NCBI WGS) accession number AGKD04000000) using BWA mem version 0.7.10-r789 (ref. 49). The thirty-two samples were sequenced to a depth of around  $18 \times$  ( $8 \times$  to  $32 \times$ ). In total, 7.6 billion out of 8.3 billion reads (92%) were properly aligned to the genome. SNPs and short indels were identified using FreeBayes (version 0.9.15-1 (ref. 50)). To filter away low-quality variants, we used the run-time parameters *-use-mapping-quality* and *-min-mapping-quality 1*, in addition to *-vcfilter -f "QUAL > 20"*. SNPs and short indels were annotated using snpEff version 4.0e (ref. 51). The snpEff annotation database was based on the CIGENE annotation version 2.0 (Lien *et al.*, submitted).

The potential effects of the missense mutations detected in *VGLL3* and *AKAP11* on protein function was assessed using the PolyPhen2 program<sup>52</sup>. Polyphen2 predicts the possible impact of amino-acid substitutions on the structure and function of proteins using physical and evolutionary comparative considerations. Owing to a lack of structural information available for these two genes, the assessment relied on evolutionary comparisons using multiple-sequence alignments.

**Identifying signatures of spatially divergent selection: iHS analysis.** Genotypic data from all individuals ( $n = 1,518$ ) were phased using Beagle 4.0 software<sup>53</sup>

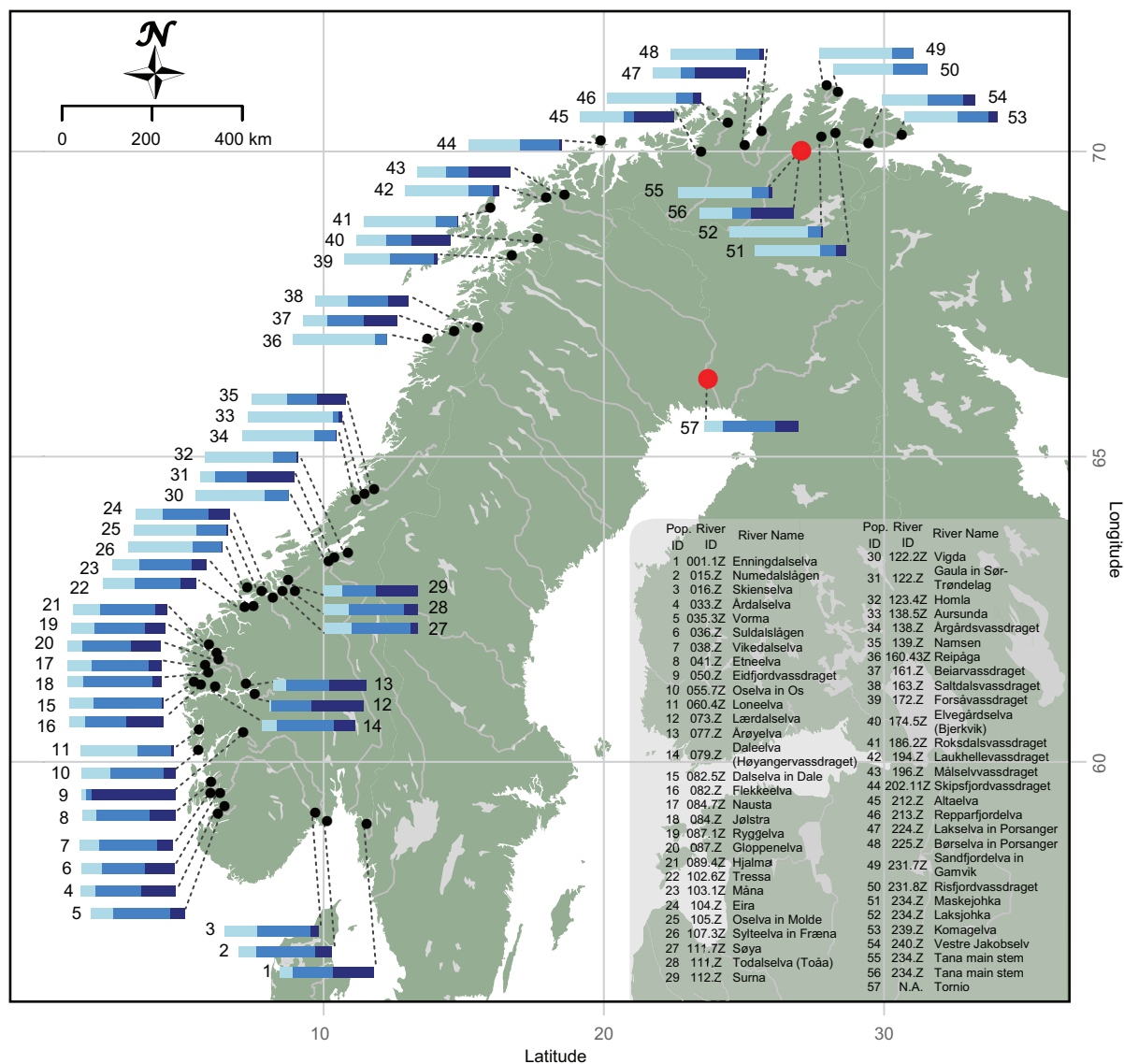
with imputation for missing genotypes (0.2% of calls) using a parameter window size = 50,000 and overlap size = 3,000 SNPs. 10, 40 and 50 iterations were parameterized for burn-in, phasing and imputation of the data, respectively, and physical distances were used as a proxy for genetic distances. We used an extended haplotype homozygosity (EHH)<sup>54</sup>-based test to detect footprints of selection, using the rehh package (version 3.1.1)<sup>55</sup>. We first computed integrated EHH scores (iHH) using the *scan\_ehh* function in rehh version 3.1.3 with default parameters. We then computed the iHS<sup>56</sup> for each population separately using the *ihh2ihs* function in rehh version 3.1.3 (frequency bin = 0.05, maf = 0.05). iHS is a metric to quantify the difference in EHH between the two alleles of a given SNP. The iHS statistic is standardized empirically to the distribution of observed iHS values over a range of SNPs with similar derived allele frequencies. The ancestral allele was initially randomly assigned for every SNP to have an even distribution of SNPs in each frequency category for standardization (Supplementary Information). The *L* allele of the *VGLL3<sub>iHS</sub>* SNP was assigned derived status in each population. Therefore, higher levels of extended homozygosity around the *L* allele compared with the *E* allele within a population are indicated by negative iHS values, and higher levels of extended homozygosity around the *E* allele compared with the *L* allele are indicated by positive iHS values.

We tested whether variation in age at maturity among populations could be maintained by selection towards an optimum age at maturity composition within each population, given gene flow is expected among the phenotypically divergent local populations sampled. Balancing selection is expected to leave similar patterns in the genome as recent positive selection, making it difficult to distinguish using haplotype-based methods such as iHS<sup>57</sup>. Additionally, such haplotype methods have reduced power to detect selection from standing genetic variation, as may arise from balancing selection<sup>58</sup>, such as spatially varying selection. However, haplotype patterns are expected to change even when selection acts on multiple haplotypes<sup>59</sup>, and haplotype-based methods, such as iHS, retain some power to detect selection so long as selective sweeps are not too soft: that is, they do not contain too many different haplotypes<sup>60</sup>. To test for divergent local selection, we employed a linear model where the iHS values were regressed over the average sea age of the populations (Extended Data Table 1). Statistical significance was assessed by comparing the regression coefficient (that is, the proportion of variation in iHS explained by age at maturity) at the locus of interest with the null distribution at the genome-wide level. For this analysis, we calculated the iHS statistic for every population with at least 16 successfully genotyped individuals (32 haplotypes from 46 populations; see also Extended Data Table 1), and used an equal number of individuals per population (by random selection of individuals if  $n > 16$ ). We assessed the effect of using 16 randomly selected individuals from the two populations in the TAN data set, and found that iHS values in the reduced set were in good agreement with the full data set (Pearson's  $r = 0.72$  and  $0.75$  for younger and older age-structured sub-populations, respectively;  $P < 10^{-16}$  for both data sets), suggesting the robustness of the iHS analysis with the sample size used (Extended Data Fig. 6e, f).

- Bourret, V. *et al.* SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Mol. Ecol.* **22**, 532–551 (2013).
- Karlsson, S., Diserud, O. H., Moen, T. & Hindar, K. A standardized method for quantifying unidirectional genetic introgression. *Ecol. Evol.* **4**, 3256–3263 (2014).
- Aykanat, T. *et al.* Low but significant genetic differentiation underlies biologically meaningful phenotypic divergence in a large Atlantic salmon population. *Mol. Ecol.* **24**, 5158–5174 (2015).
- Johnston, S. E. *et al.* Genome-wide SNP analysis reveals a genetic basis for sea-age variation in a wild population of Atlantic salmon (*Salmo salar*). *Mol. Ecol.* **23**, 3452–3468 (2014).
- Yano, A. *et al.* The sexually dimorphic on the Y-chromosome gene (sdY) is a conserved male-specific Y-chromosome sequence in many salmonids. *Evol. Appl.* **6**, 486–496 (2013).
- Friedland, K. D. & Haas, R. E. Marine post-smolt growth and age at maturity of Atlantic salmon. *J. Fish Biol.* **48**, 1–15 (1996).
- Fisher, J. P. & Pearcy, W. G. Spacing of scale circuli versus growth-rate in young Coho salmon. *Fish Bull.* **88**, 637–643 (1990).
- ICES. *Report of the Workshop on Age Determination of Salmon (WKADS)*. Report CM 2011/ACOM:44 (ICES, 2011).
- Einum, S., Thorstad, E. B. & Næsje, T. F. Growth rate correlations across life-stages in female Atlantic salmon. *J. Fish Biol.* **60**, 780–784 (2002).
- Jonsson, N. & Jonsson, B. Sea growth, smolt age and age at sexual maturation in Atlantic salmon. *J. Fish Biol.* **71**, 245–252 (2007).
- Gjedrem, T., Gjøs, H. M. & Gjerde, B. Genetic origin of Norwegian farmed Atlantic salmon. *Aquaculture* **98**, 41–50 (1991).
- GenABEL Project Developers. GenABEL: genome-wide SNP association analysis. R package version 1.8-0 (2013).
- R Core Team. R: a language and environment for statistical computing (R Foundation for Statistical Computing, 2014).

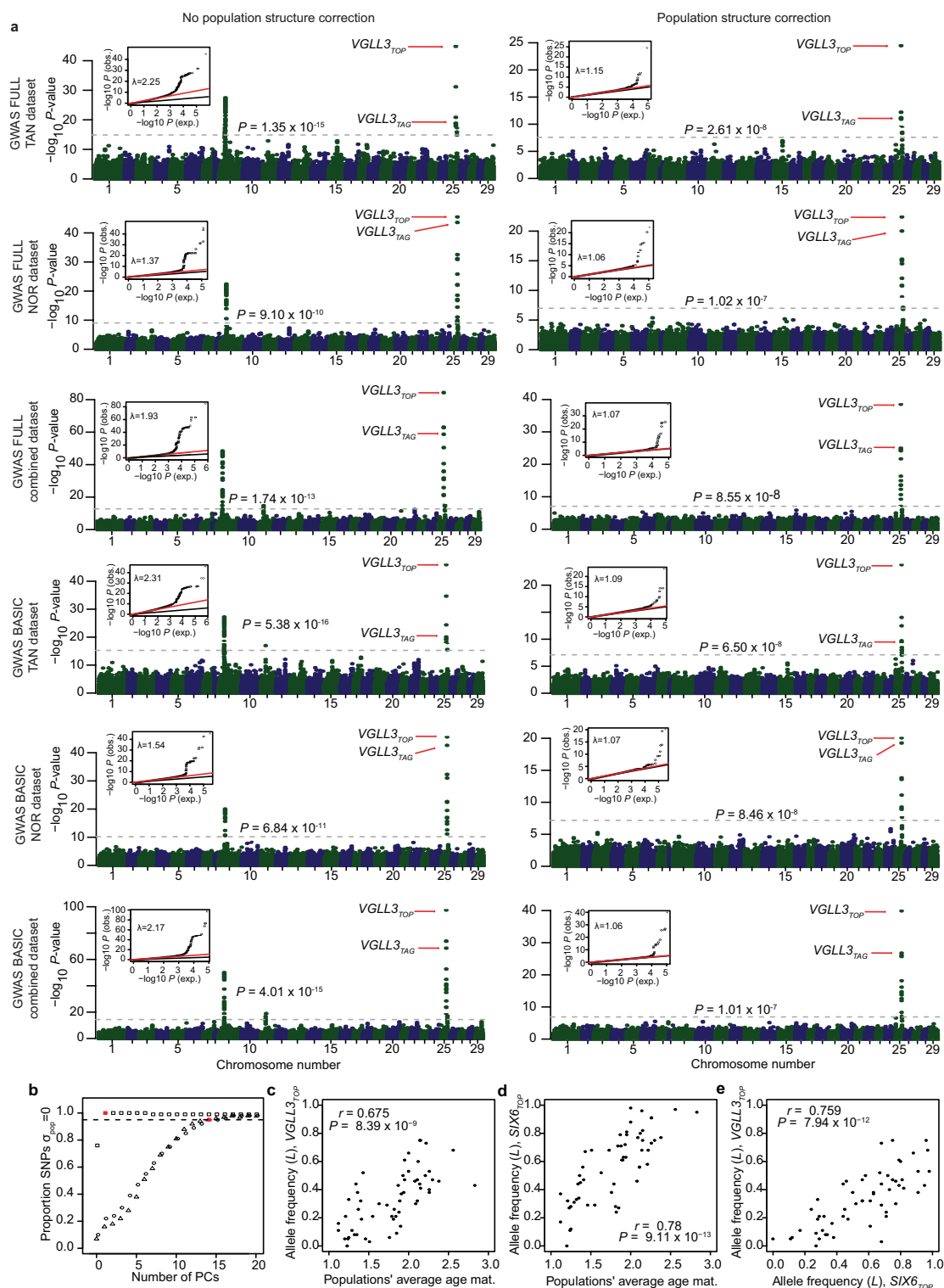
44. Christensen, R. H. B. ordinal - regression models for ordinal data. R package version 2015.1-21 (2015).
45. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* **96**, 329–339 (2015).
46. Bonhomme, M. *et al.* Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* **186**, 241–262 (2010).
47. Lotterhos, K. E. & Whitlock, M. C. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol. Ecol.* **23**, 2178–2192 (2014).
48. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, v067.i01 (2015).
49. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
50. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <http://arXiv.org/abs/1207.3907> (2012).
51. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain *w<sup>1118</sup>*; *iso-2*; *iso-3*. *Fly* **6**, 80–92 (2012).
52. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
53. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
54. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
55. Gautier, M. & Vitalis, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176–1177 (2012).
56. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
57. Fijarczyk, A. & Babik, W. Detecting balancing selection in genomes: limits and prospects. *Mol. Ecol.* **24**, 3529–3545 (2015).
58. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**, 659–669 (2013).
59. Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.* **31**, 1275–1291 (2014).
60. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet.* **11**, e1005004 (2015).





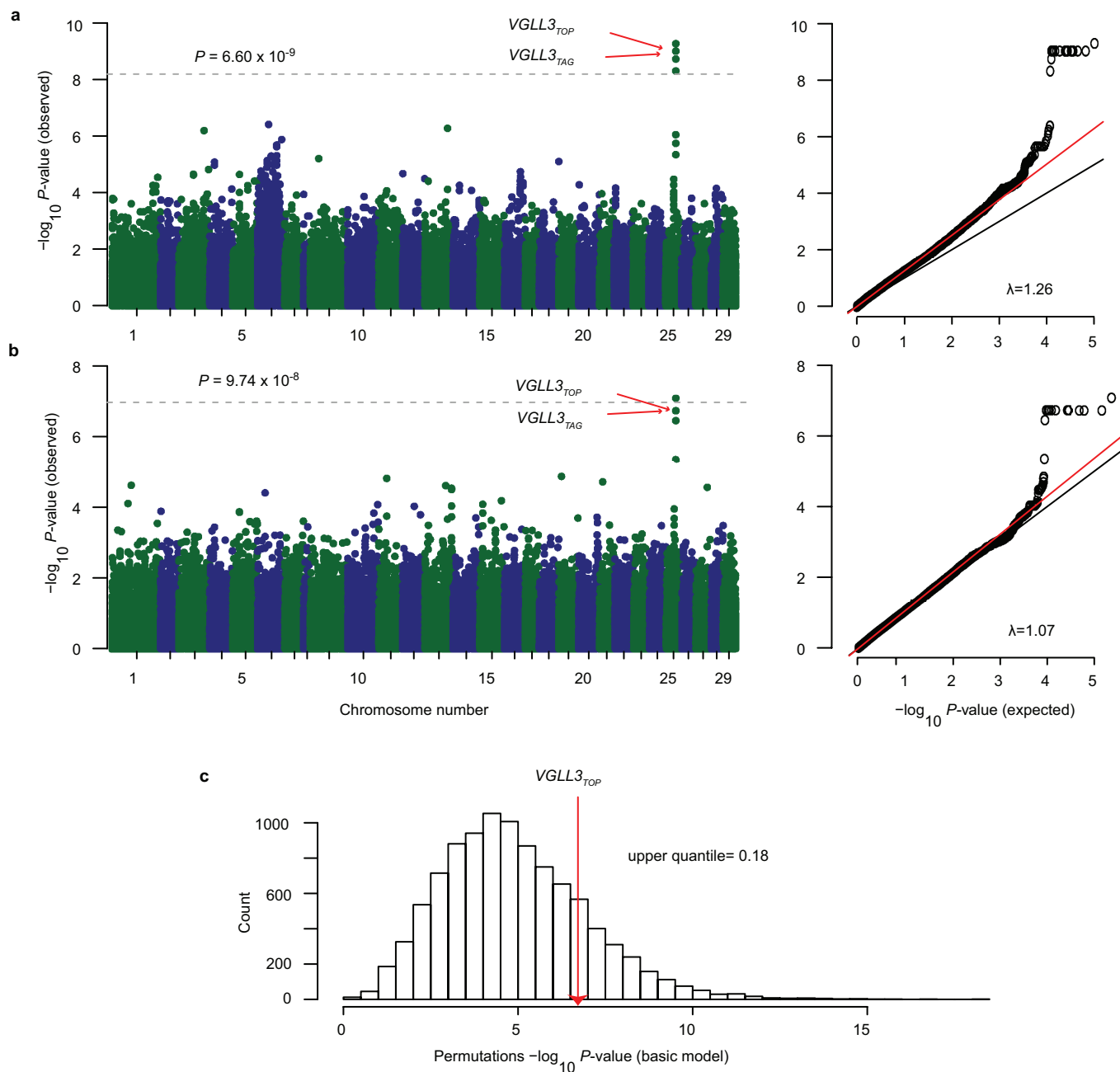
**Extended Data Figure 1 | Map of study populations.** Bars indicate the proportion of individuals maturing after 1 (light blue), 2 (medium blue) or  $\geq 3$  years (dark blue) at sea; 1–54, NOR data set; 55–56, TAN; 57, BAL (Extended Data Table 1). Data for lake and river coordinates were obtained

from European Environmental Agency (under a Creative Commons Attribution 4 License) and the Norwegian Water Resources and Energy Directorate.



**Extended Data Figure 2 | GWAS analyses for the TAN ( $n=463$ ), NOR ( $n=941$ ) and combined ( $n=1,404$ ) data sets. a**, Manhattan and quantile–quantile plots of the GWAS for age at maturity in Atlantic salmon before (left) and after (right) correction for population structure. The first three rows are models including phenotypic covariates (that is, the FULL model), and the next three rows are models without phenotypic covariates (that is, the BASIC model). The  $y$  axis shows the association statistic ( $-\log_{10}(P$  values)) for each SNP ordered by chromosome and position ( $x$  axis). The genome-wide statistical significance adjusted for multiple comparisons and genomic inflation is indicated by a horizontal dashed line. The  $VGLL3_{TOP}$  (the SNP with the highest association with age at maturity) and  $VGLL3_{TAG}$  (the SNP strongest linkage disequilibrium

with the missense mutations in the  $VGLL3$  gene) SNPs are shown with red arrows. QQ plots showing the deviation of  $P$  values (red line) from the null expectation (black line) are in the insets. **b**, Proportion of SNPs showing no evidence of significant population structure ( $H_{null}$ : Akaike information criterion  $< -2$ ) as a function of the number of principal components included in the model, for TAN (squares), NOR (circles) and the combined data set (TAN + NOR; triangles). The numbers of principal components used in population corrected models are marked with red. **c**, Relationship between population average age at maturity and allele frequency at the  $VGLL3_{TOP}$  SNP and **(d)**  $SIX6_{TOP}$  SNP. **e**, Relationship between the  $VGLL3_{TOP}$  SNP and the  $SIX6_{TOP}$  SNP allele frequencies.

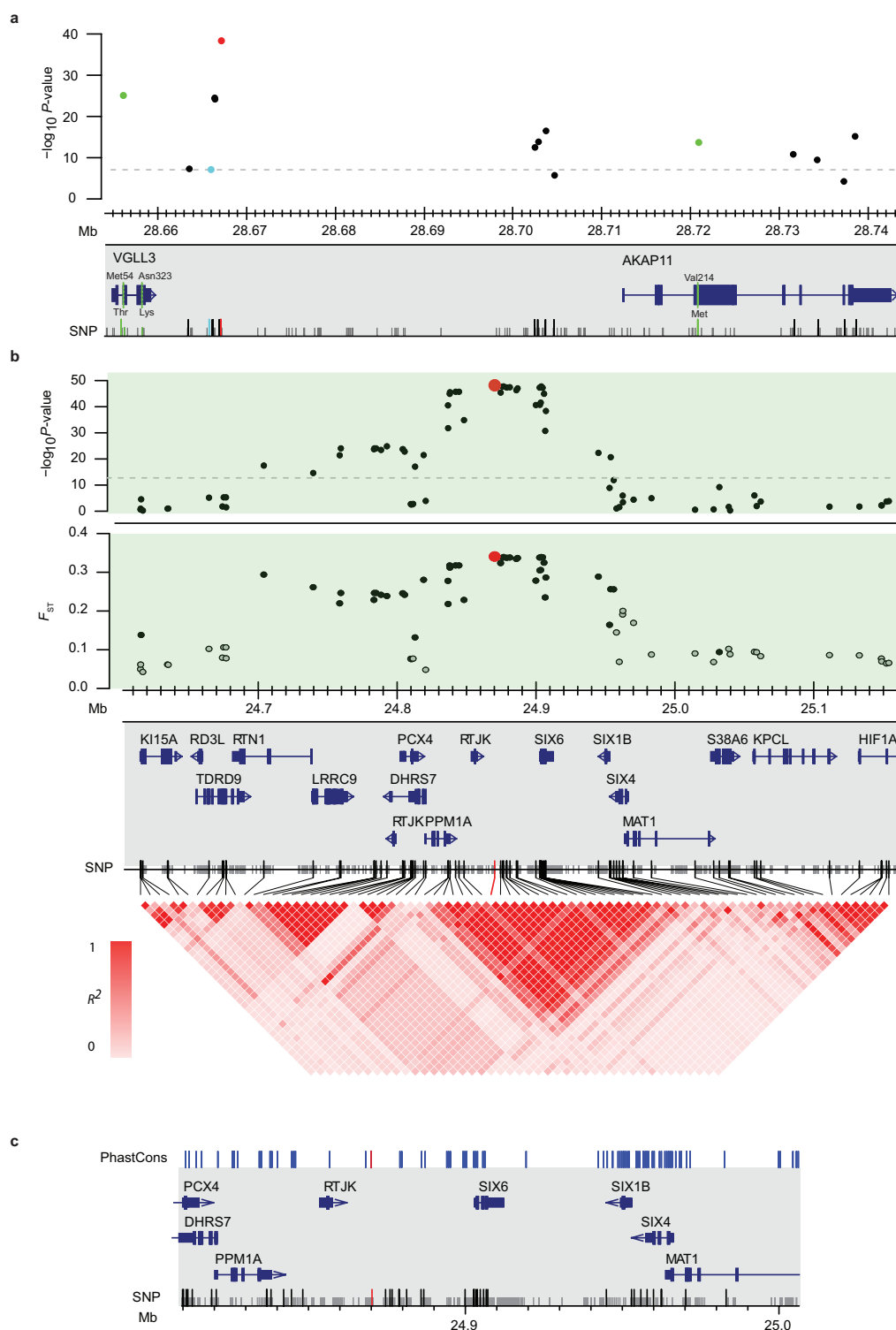


### Extended Data Figure 3 | GWAS analyses for the BAL data set.

Manhattan plots and quantile–quantile plots of the GWAS for age at maturity in the BAL data set ( $n = 114$ ), (a) before and (b) after correction for population structure. The y axis shows the association statistic ( $-\log_{10}(P\text{ values})$ ) for each SNP ordered by chromosome and position (x axis). The genome-wide statistical significance adjusted for multiple comparisons and genomic inflation is indicated by a horizontal dashed line. The  $VGLL3_{TOP}$  and  $VGLL3_{TAG}$  SNPs are shown with red arrows.

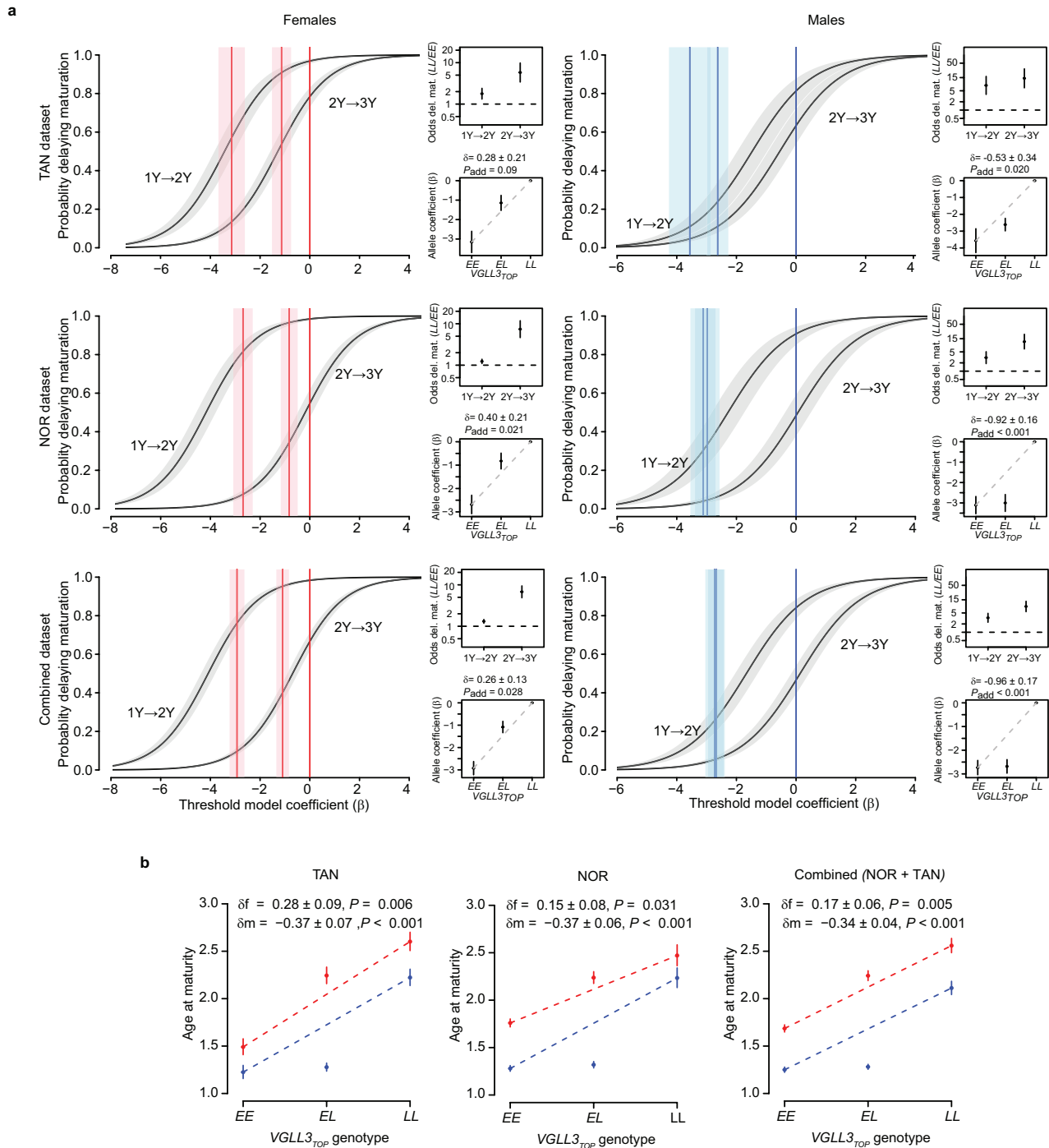
The QQ plot shows the deviation of  $P$  values (red line) from the null expectation (black line). c, Distribution of association statistics for the  $VGLL3_{TOP}$  SNP in 100,000 bootstrapped replicates with resampling, using the TAN + NOR data set combined ( $n = 1,404$ ). An equivalent sampling design to the BAL data set ( $n = 114$  and the same age at maturity structure; see Supplementary Table 1) was used in the resampling. The red arrow indicates the  $P$  value of the  $VGLL3_{TOP}$  SNP in the BAL data set.





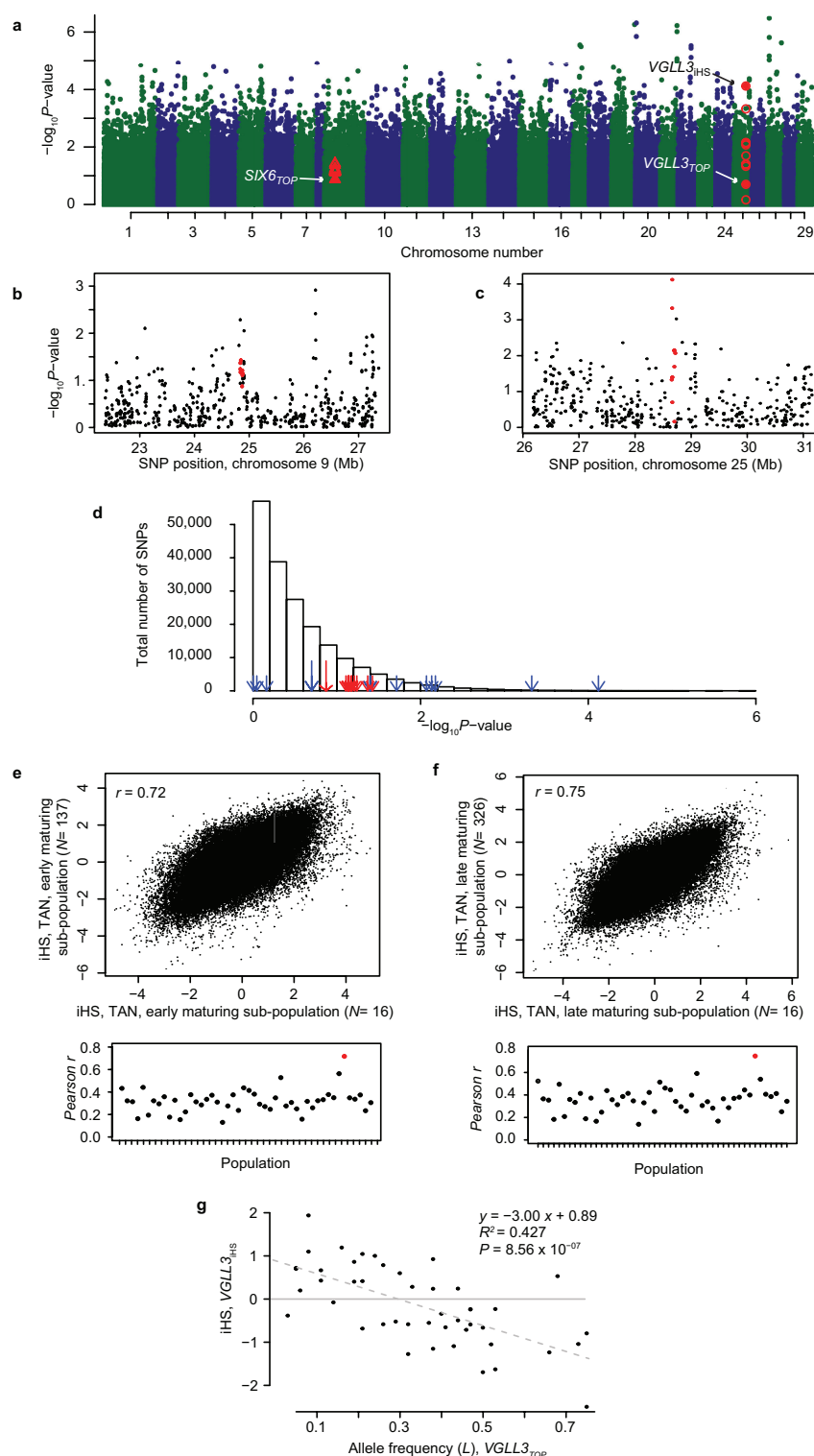
**Extended Data Figure 4 | Gene model diagrams detailing regions around the *VGLL3*<sub>TOP</sub> and *SIX6*<sub>TOP</sub> loci. a**, Gene models and genomic positions of the two genes in the genome region on chromosome 25 significantly associated with age at maturity. Missense SNPs identified by re-sequencing within the genes are indicated in green. Amino acids indicated above and below the gene model were associated with the late (L) and early (E) maturation alleles, respectively. Longer tick marks show custom 220K Affymetrix axion array SNPs, and shorter tick marks indicate re-sequencing variants. Notable SNPs are colour coded with red (*VGLL3*<sub>TOP</sub>), blue (*VGLL3*<sub>IHS</sub>) and green (the SNP tagging missense mutations in *VGLL3* and the *AKAP11* missense SNP). Note that missense variants on *VGLL3* were identified by whole genome sequencing. The array SNP in tightest linkage disequilibrium with the *VGLL3* missense variants identified by re-sequencing is 306 and 2,356 base pairs upstream

( $R^2 = 1$  and 0.71, respectively). **b**, Gene model and linkage disequilibrium plots of an ~0.5 Mb region on chromosome 9 where a significant GWAS signal was observed before correction for population structure. The association plot shown is before correction for population structure, using the combined data set (TAN + NOR). The *SIX6*<sub>TOP</sub> locus is shown in red. Shorter tick marks in the SNP axis indicate re-sequencing variants.  $F_{ST}$  estimates for SNPs in the region are also shown (lower graph). Closed circles indicate SNPs significantly diverged from null (neutral) expectations (FLK  $F_{ST}$  outlier test, 99.5% quantile of the null distribution, (56 populations, total  $n = 1,404$ ). **c**, Conserved elements (PhastCons) of the 200 kb region around the *SIX6* gene showing the predicted forebrain distal regulatory element (red tick mark) that is located close to the *SIX6*<sub>TOP</sub> SNP. One re-sequenced variant in strong linkage disequilibrium with the *SIX6*<sub>TOP</sub> SNP was located in this region.



**Extended Data Figure 5 | Details of modelling the genetic architecture of age at maturity.** **a**, Threshold logistic models explaining variation in age at maturity in relation to the VGLL3<sub>TOP</sub> SNP in the TAN ( $n = 220$  females, 243 males), NOR ( $n = 473$  females, 468 males) and the combined ( $n = 693$  females, 711 males) data sets for females (left panels) and males (right panels). Shaded grey areas around the logistic curves indicate one standard error of the threshold coefficients, and shaded red and blue areas indicate one standard error around genotype coefficients for females and males, respectively. The y axis depicts the probability of delaying maturation from one maturity age class to the next. LL genotypes were centred to zero (intercept) and had no standard error because of the rank deficiency of the model (that is, threshold degrees of freedom is prioritized in the model). Threshold coefficients are sex independent, which was the optimal model explaining the data (see Extended Data Table 2 and Supplementary information 3). Small insets to the right of each logistic curve depict the odds of delaying maturation for the LL genotype in relation to the EE genotype (median, 50% parametric sampling

quantile) and the degree of partial dominance (median, 50% parametric sampling quantile) on the unobserved liability scale (that is, the x axis in the logistic curves). The dominance estimates ( $\delta$ ) given above each panel are scaled to  $[-1, 1]$  range ( $\delta = (2\beta_{EL} + (\beta_{LL} - \beta_{EE})) / (|\beta_{LL} - \beta_{EE}|)$ ), where negative and positive values indicate an EE-like, and LL-like, expression of the phenotype (that is, delayed maturation), respectively.  $P$  values in the upper insets show the significance of the model deviating from additivity ( $P_{\text{add}}$ , 10,000 parametric permutations). The difference in dominance between females and males is highly significant for all data sets ( $P = 0.0082$  for TAN, and  $P < 0.001$  for NOR and the combined data sets.).  $P$  values for all odds of delaying maturation are significant ( $P < 0.001$ , 100,000 parametric permutations). **b**, Predicted mean and 50% sampling quantiles (10,000 parametric permutations) of age at maturity using the logit transformation model. The y axis is log scaled.  $P_{\text{add}}$  values in the insets shows significance of the model deviating from additivity (10,000 parametric permutations).



#### Extended Data Figure 6 | Haplotype length analysis summary.

**a**, Manhattan plot of each SNP in the study showing the  $P$  values of the correlation between population *iHS* values (46 populations, 32 haplotypes per population) and the average age at maturity. Ten SNPs flanking the *VGLL3*<sub>TOP</sub> and *SIX6*<sub>TOP</sub> SNPs are marked with red circles and triangles, respectively. **b**, **c**, Same as **a** but showing a 5 Mb magnified view of the (**b**) *VGLL3* and (**c**) *SIX6* regions. **d**, Histogram showing the statistic distribution of the association between *iHS* and average age at maturity for all SNPs analysed in the study. Ten SNPs around the *VGLL3*<sub>TOP</sub> and *SIX6*<sub>TOP</sub> SNPs are marked with blue and red arrows, respectively, where longer arrow tails show the *VGLL3*<sub>TOP</sub> and *SIX6*<sub>TOP</sub> SNPs. **e**, **f**, *iHS* concordance (Pearson's  $r$ ) in the TAN data set between the

reduced ( $n = 16$ ) and full data sets for (**e**) a sub-population (55) with lower average age at maturity ( $n = 137$ ) and (**f**) a sub-population (56) with higher average age at maturity ( $n = 326$ ). Each point shows a single SNP. The lower panel shows the concordance (Pearson's  $r$ ) of the TAN full data sets to all populations ( $n = 46$ ) included in the *iHS* analysis. The self-concordance, as in the upper panel, is indicated with red. **g**, Relationship between population *iHS* score and *VGLL3*<sub>TOP</sub> allele frequency. *iHS* = 0 (no haplotype length difference) is marked with a horizontal grey line. Positive *iHS* values indicate longer haplotype blocks, and therefore stronger selection, around the *E* allele in a population relative to the *L* allele, and vice versa for negative *iHS* values.

**Extended Data Table 1 | Geographic and life-history details of Atlantic salmon populations included in this study, with sample sizes and genetic data of key SNPs for each population**

Pop. ID*	River code†	River name	Phylo‡	Population details								Summary of study samples						
				Coordinates	N <sub>T</sub> §	% Maiden¶	Mat age	Proportion mat age (%)				L(cm, mean)	N <sup>  </sup>	Mat age <sup>#</sup>	Year (20xx)	H <sub>O</sub>	iHS*	Allele freq. (L)
				Lat(N) Lon(E)				1 yr.	2 yr.	3 yr.	>3 yr.							
<b>NOR</b>																		
1	001.1Z	Enningdalselva	ATL	58.98 11.47	824	92.4	2.30	14.4	42.6	42.3	0.8	78.7	16(5)	1.94	12	0.56	-0.242	0.75 0.47
2	015.Z	Numedalslågen	ATL	59.03 10.06	1483	95.5	2.00	18.7	63.9	16.3	1.1	77.7	16(7)	1.81	12	0.69	-1.237	0.88 0.66
3	016.Z	Skienelva	ATL	59.13 9.63	1024	91.6	1.73	34.6	57.3	7.8	0.2	69.2	17(12)	1.82	12	0.35	-0.523	0.68 0.29
4	033.Z	Årdalselva	ATL	59.14 6.17	1102	94.7	2.21	15.2	48.9	34.4	1.6	78.0	18(10)	2.00	11	0.56	-1.697	0.58 0.50
5	035.3Z	Vorma	ATL	59.27 6.33	2534	98.0	1.90	24.3	60.2	14.8	0.6	70.6	18(9)	1.94	11	0.06	-0.078	0.61 0.14
6	036.Z	Suldalslågen	ATL	59.48 6.25	2070	98.3	2.11	22.0	46.4	28.9	2.7	78.5	17(9)	1.82	11	0.53	-0.585	0.56 0.32
7	038.Z	Vikedalselva	ATL	59.49 5.90	193	94.3	1.95	21.3	62.6	16.1	0	72.0	14(6)	1.86	12	0.79		0.79 0.61
8	041.Z	Etnelva	ATL	59.67 5.93	1194	91.5	2.14	15.2	57.0	26.4	1.4	76.3	17(7)	2.24	12	0.29	-0.495	0.68 0.44
9	050.Z	Eidfjordvassdraget	ATL	60.47 7.07	23	100	2.83	5.3	5.3	84.2	5.3	98.3	21(15)	2.86	11-12	0.29	-1.093	0.95 0.43
10	055.7Z	Oselva in Os	ATL	60.18 5.47	1029	99.1	1.82	30.2	56.9	12.6	0.3	68.9	17(9)	1.88	11	0.29	-0.585	0.24 0.26
11	060.4Z	Loneelva	ATL	60.53 5.49	428	91.6	1.41	61.6	36.2	2.2	0	60.1	16(4)	1.56	12	0.25	0.861	0.47 0.19
12	073.Z	Lærdalselva	ATL	61.10 7.48	149	97.3	2.56	1.5	43.4	50.0	5.1	92.0	17(15)	3.06	7	0.53	0.529	0.97 0.68
13	077.Z	Arøyelva	ATL	61.27 7.17	180	100	2.26	13.7	46.4	38.1	1.8	82.7	16(5)	2.19	12	0.44	-1.630	0.91 0.53
14	079.Z	Daleelva Høyangervassdraget	ATL	61.22 6.07	407	90.9	2.08	16.4	60.7	22.1	0.8	75.0	13(7)	2.15	12	0.46		0.46 0.46
15	082.5Z	Dalselva in Dale	ATL	61.36 5.40	66	100	1.76	26.2	72.3	1.5	0	63.8	17(5)	1.24	13	0.12	0.198	0.38 0.06
16	082.Z	Flekkeelva	ATL	61.31 5.35	1816	97.5	2.22	17.1	43.8	37.7	1.4	76.3	20(14)	2.35	11	0.45	-1.043	0.68 0.73
17	084.7Z	Nausta	ATL	61.51 5.72	498	88.6	1.85	25.9	60.8	13.1	0.3	68.3	11(7)	1.82	12	0.18		0.27 0.09
18	084.Z	Jølstra	ATL	61.46 5.83	133	97.7	1.91	17.6	73.1	9.2	0	73.2	19(11)	1.89	13	0.42	-0.553	0.63 0.37
19	087.1Z	Ryggeelva	ATL	61.78 6.13	253	100	1.93	24.3	53.9	21.8	0	73.5	18(9)	1.89	11	0.56	-0.664	0.72 0.50
20	087.Z	Gloppenelva	ATL	61.77 6.20	1215	97.1	2.15	16.1	52.2	29.8	1.9	77.1	26(16)	2.62	11	0.50	-2.499	0.75 0.75
21	089.4Z	Hjalma	ATL	61.91 5.85	206	98.5	1.84	28.7	58.7	12.6	0	67.7	7(2)	2.00	11	0.43		0.71 0.21
22	102.6Z	Tressa	ATL	62.52 7.13	156	94.2	1.81	34.5	48.9	16.5	0	65.4	18(10)	1.44	12	0.39	0.402	0.28 0.19
23	103.1Z	Måna	ATL	62.54 7.44	166	95.8	1.89	28.3	56.5	15.2	0	69.8	14(10)	1.79	12	0.50		0.61 0.32
24	104.Z	Eira	ATL	62.68 8.12	1141	95.8	1.95	28.6	49.1	21.1	1.2	73.8	18(10)	2.39	12	0.39	-0.591	0.81 0.47
25	105.Z	Oselva in Molde	ATL	62.79 7.72	502	95.0	1.35	66.2	32.5	1.3	0	57.2	19(9)	1.47	12	0.16	1.098	0.34 0.08
26	107.3Z	Sylteelva in Fræna	ATL	62.84 7.21	820	96.0	1.32	68.9	30.3	0.8	0	56.0	18(7)	1.50	12	0.17	1.937	0.33 0.08
27	111.7Z	Søya	ATL	62.89 8.54	68	85.3	1.79	29.8	63.2	7.0	0	66.3	17(11)	1.82	12	0.29	0.414	0.44 0.21
28	111.Z	Todalselva (Toåa)	ATL	62.82 8.70	98	96.9	1.88	26.7	58.9	13.3	1.1	72.2	17(10)	2.00	12	0.24	1.000	0.71 0.24
29	112.Z	Suma	ATL	62.97 8.67	1198	95.2	2.26	20.1	35.2	42.2	2.5	79.0	20(8)	1.9	13	0.40	-0.343	0.73 0.40
30	122.2Z	Vigda	ATL	63.31 10.18	83	92.8	1.25	74.7	25.3	0	0	51.2	17(15)	1.24	10	0.12	NA	0.12 0.06
31	122.Z	Gaula in Sør-Trøndelag	ATL	63.34 10.24	1218	91.9	2.38	15.0	34.5	48.0	2.5	83.7	24(14)	2.67	13	0.50	-0.714	0.77 0.46
32	123.4Z	Homla	ATL	63.41 10.80	113	94.7	1.29	72.9	25.2	1.9	0	56.2	18(11)	1.39	11	0.22	0.665	0.31 0.11
33	138.5Z	Aursunda	ATL	64.37 11.39	124	91.9	1.12	91.2	5.3	3.5	0	47.8	18(11)	1.00	11	0.22	0.427	0.17 0.11
34	138.Z	Årgårdsvassdraget	ATL	64.31 11.22	1335	94.0	1.23	77.5	22.3	0.2	0	53.4	13(7)	1.54	12	0.00		0.27 0.00
35	139.Z	Namsen	ATL	64.46 11.52	1308	93.3	1.94	37.5	31.6	29.9	1.0	71.9	16(4)	1.63	12	0.63	0.924	0.63 0.38
36	160.43Z	Reipåga	ATL	66.91 13.63	38	86.8	1.12	87.1	12.9	0	0	52.5	19(8)	1.42	11-13	0.21	1.192	0.37 0.16
37	161.Z	Beiarvassdraget	ATL	67.03 14.58	1561	92.1	2.09	26.0	38.6	34.4	1.0	77.0	15(9)	2.07	12	0.60		0.77 0.30
38	163.Z	Saltålvassdraget	ATL	67.10 15.42	983	94.6	1.87	35.5	42.6	21.0	0.9	74.2	7(3)	2.71	12	0.29		0.79 0.43
39	172.Z	Forsåvassdraget	ATL	68.27 16.63	117	87.2	1.55	49.0	47.1	3.9	0	63.3	17(10)	1.47	12	0.41	1.044	0.29 0.21
40	174.5Z	Elvegårdselva (Bjerkvik)	ATL	68.55 17.56	40	92.5	2.11	32.4	26.5	41.2	0	76.8	16(7)	1.75	11-12	0.31	-0.236	0.56 0.47
41	186.2Z	Roksdalsvassdraget	ATL	69.05 15.87	753	94.3	1.24	76.7	22.7	0.6	0	55.2	19(11)	1.37	12	0.11	NA	0.11 0.05
42	194.Z	Laukhellevassdraget	ATL	69.23 17.86	138	93.5	1.37	67.7	25.8	6.5	0	58.9	19(7)	1.53	12	0.42	0.785	0.45 0.26
43	196.Z	Målselvassdraget	BW	69.27 18.51	591	94.4	2.15	31.5	23.5	43.9	1.0	78.6	17(5)	1.47	11-12	0.47	-0.656	0.82 0.41
44	202.11Z	Skipsfjordvassdraget	ATL	70.16 19.80	137	95.6	1.46	55.8	41.7	2.5	0	57.2	19(7)	1.32	12	0.11	0.698	0.29 0.05
45	212.Z	Altaelva	BW	69.97 23.37	2047	97.3	2.00	46.7	10.6	38.3	4.4	76.3	20(5)	2.00	12	0.55	-0.233	0.98 0.53
46	213.Z	Repparfjordelva	BW	70.45 24.32	3932	97.5	1.36	73.1	18.1	8.5	0.3	61.2	17(6)	1.94	12	0.53	0.237	0.50 0.38
47	224.Z	Lakselva in Porsanger	BW	70.08 24.92	361	94.5	2.26	29.9	15.3	48.1	6.7	84.6	16(7)	2.06	12	0.50	-1.153	0.91 0.38
48	225.Z	Børselva in Porsanger	BW	70.31 25.52	287	95.5	1.34	70.3	25.3	4.0	0.4	61.5	17(5)	1.71	11	0.41	0.241	0.44 0.44
49	231.7Z	Sandfjordelva in Gamvik	BW	71.05 28.05	149	94.0	1.21	77.5	22.5	0	0	57.9	17(9)	1.29	12	0.41	-0.685	0.32 0.21
50	231.8Z	Risfjordvassdraget	BW	70.98 28.17	46	93.5	1.40	63.4	36.6	0	0	58.6	17(5)	1.41	11	0.65	-1.276	0.56 0.32
51	234.Z	Maskelohka	BW	70.28 28.15	327	98.2	1.43	70.3	17.4	8.0	2.4	74.8	30(15)	1.90	03-10	0.50	-1.052	0.67 0.52
52	234.Z	Lakselohka	BW	70.06 27.55	126	100	1.20	84.9	13.5	0	1.6	57.4	21(12)	1.52	03-10	0.10	0.721	0.00 0.05
53	239.Z	Komagelva	BW	70.24 30.52	1029	93.6	1.51	57.4	32.7	9.9	0	63.6	17(11)	1.76	12	0.06	-0.385	0.68 0.03
54	240.Z	Vestre Jakobselv	BW	70.11 29.33	2038	97.1	1.63	49.9	37.3	12.4	0.4	68.3	23(10)	1.65	13	0.43	0.598	0.52 0.30
<b>TAN</b>																		
55-56 <sup>‡‡</sup>	234.Z	Tana main stem	BW	70.47 28.25	86317	93.7	1.60	61.5	18.1	19.1	1.3	72.0	463(220)	2.15	01-03			
55							1.25	47.6	10.5	2.2	0		137(71)	1.36		0.44	0.282	0.28 0.33
56							2.14	13.9	7.6	16.9	1.3		326(149)	2.49		0.39	-0.796	0.96 0.75
<b>BAL</b>																		
57		Tomio	BAL	65.81 24.16	4822	93.8	2.12	16.4	56.2	26.9	0.6	83.4	114(12)	1.94	05-08	0.50		0.80 0.60

\*The unique population ID used in this study (see also Extended Data Fig. 1).

†Unique code for Norwegian rivers.

‡Phylogeographic lineage as in ref. 31. ATL, Atlantic; BW, Barents/White; BAL, Baltic.

§The total number of samples used to infer population age-structure data. Years of collection were 2006–2014 for all populations except 1971–2009 for 55–56, 2000–2013 for 57, 1977–2009 for 51, 1985–2009 for 52, 2006–2012 for 9, 2006–2007 for 12, 2006–2013 for 15.

|| Percentage of maiden fish (first time spawners) in the population data.

¶Number of individuals analysed with the SNP array (number of females in parentheses). Number of individuals genome sequenced was  $n = 3$  for populations 17, 18, 34, 35, 45, 46, and  $n = 14$  for 55 and 56.

#The average sea age at maturity of individuals analysed in the GWAS.

★Population-specific iHS. iHS was not calculated for some populations because of either low sample size (blank) or low maf ( $< 0.05$ ; NA).\*\*Frequency of the *SIX6<sub>TOP</sub>* allele associated with populations with older age at maturity.††Frequency of the *VGLL3<sub>TOP</sub>* allele associated with older age at maturity.‡‡Populations 55 and 56 coexist sympatrically in the main stem Tana River and have younger and older age structures, respectively<sup>33</sup>. Therefore, sea age proportions of these sub-populations are not directly available, and were extrapolated using weighted proportions of sea age classes assigned to each sub-population in ref. 33.



**Extended Data Table 2 | Quality of various genetic architecture models explaining sea age at maturity at the *VGLL3<sub>TOP</sub>* locus**

model number	model name	model description	AIC <i>TAN</i>	$\Delta$ AIC <i>TAN</i>	AIC <i>NOR</i>	$\Delta$ AIC <i>NOR</i>	AIC combined	$\Delta$ AIC combined
1	Additive	Allelic effects are linear: e.g. <i>pp</i> , <i>pq</i> , <i>qq</i> coded as 0,1,2	604.95	22.58	1215.50	62.53	1814.01	96.43
2	Dominance type I	Dominance model to first allele: e.g. <i>pp</i> , <i>pq</i> , <i>qq</i> coded as 0,0,2	674.45	92.07	1268.42	115.45	1916.47	198.89
3	Dominance type II	Dominance model to second allele: e.g. <i>pp</i> , <i>pq</i> , <i>qq</i> coded as 0,2,2	618.52	36.14	1223.93	70.97	1824.46	106.88
4	Dominance - sex type I	Sex specific full dominance model type I: e.g. <i>pp</i> , <i>pq</i> , <i>qq</i> coded as 0,2,2 for males, and 0,0,2 for females.	602.20	19.83	1200.37	47.40	1792.90	75.32
5	Dominance - sex type II	Sex specific full dominance model type II: e.g. <i>pp</i> , <i>pq</i> , <i>qq</i> coded as 0,0,2 for males, and 0,2,2 for females.	683.72	101.35	1289.16	136.20	1941.96	224.37
6	Additive-dominance	Partial dominance is modelled by a genotype model (i.e. all genotypes modelled independently).	604.69	22.32	1204.57	51.60	1796.08	78.49
7	Additive-dominance with sex, type I	Sex specific partial dominance is modelled by coding genotypes independently and interacting with sex.	594.81	12.43	1171.84	18.88	1755.59	38.01
8	Additive-dominance with sex, type II	Sex dependent partial dominance is modelled by coding genotypes independently and sex interaction, plus sea age threshold levels sex specifically modelled.	582.37	0.00	1152.97	0.00	1717.58	0.00

The FULL model (that is, with phenotypic covariates) with population structure was employed to TAN ( $n = 463$ ), NOR ( $n = 941$ ) and the combined ( $n = 1,404$ ) data sets.

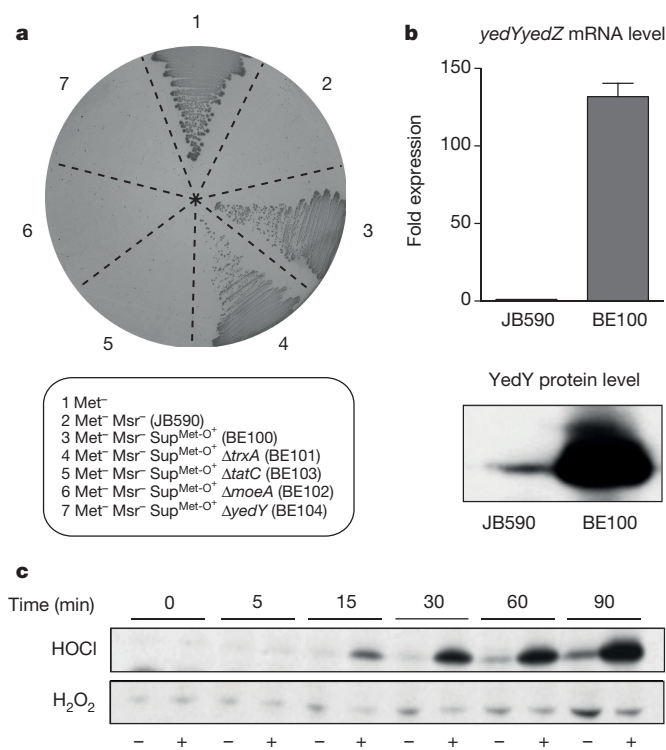
# Repairing oxidized proteins in the bacterial envelope using respiratory chain electrons

Alexandra Gennaris<sup>1,2,3,\*</sup>, Benjamin Ezraty<sup>4,\*</sup>, Camille Henry<sup>4</sup>, Rym Agrebi<sup>1,2,3</sup>, Alexandra Vergnes<sup>4</sup>, Emmanuel Oheix<sup>5</sup>, Julia Bos<sup>4,†</sup>, Pauline Leverrier<sup>1,2,3</sup>, Leon Espinosa<sup>4</sup>, Joanna Szewczyk<sup>1,2,3</sup>, Didier Vertommen<sup>2</sup>, Olga Iranzo<sup>5</sup>, Jean-François Collet<sup>1,2,3</sup> & Frédéric Barras<sup>4</sup>

The reactive species of oxygen and chlorine damage cellular components, potentially leading to cell death. In proteins, the sulfur-containing amino acid methionine is converted to methionine sulfoxide, which can cause a loss of biological activity. To rescue proteins with methionine sulfoxide residues, living cells express methionine sulfoxide reductases (Msrs) in most subcellular compartments, including the cytosol, mitochondria and chloroplasts<sup>1–3</sup>. Here we report the identification of an enzymatic system, MsrPQ, repairing proteins containing methionine sulfoxide in the bacterial cell envelope, a compartment particularly exposed to the reactive species of oxygen and chlorine generated by the host defence mechanisms. MsrP, a molybdo-enzyme, and MsrQ, a haem-binding membrane protein, are widely conserved throughout Gram-negative bacteria, including major human pathogens. MsrPQ synthesis is induced by hypochlorous acid, a powerful antimicrobial released by neutrophils. Consistently, MsrPQ is essential for the maintenance of envelope integrity under bleach stress, rescuing a wide series of structurally unrelated periplasmic proteins from methionine oxidation, including the primary periplasmic chaperone SurA. For this activity, MsrPQ uses electrons from the respiratory chain, which represents a novel mechanism to import reducing equivalents into the bacterial cell envelope. A remarkable feature of MsrPQ is its capacity to reduce both rectus (R-) and sinister (S-) diastereoisomers of methionine sulfoxide, making this oxidoreductase complex functionally different from previously identified Msrs. The discovery that a large class of bacteria contain a single, non-stereospecific enzymatic complex fully protecting methionine residues from oxidation should prompt a search for similar systems in eukaryotic subcellular oxidizing compartments, including the endoplasmic reticulum.

The fact that no Msr had been identified in the cell envelope of important human pathogens, including *Escherichia coli* and *Pseudomonas aeruginosa*, was surprising as this compartment is particularly exposed to the oxidizing compounds present in the environment. We postulated that such a methionine sulfoxide (Met-O) reducing system had remained unidentified, and applied a genetic approach to uncover it, using *E. coli* as a model. We first constructed an *E. coli* Met auxotroph mutant lacking all cytoplasmic Msrs and found this strain (JB590) to be unable to use Met-O as the only Met source (Fig. 1a). We then searched for suppressor mutations conferring Met-O reducing capacity to JB590, which led to the isolation of strain BE100 (Fig. 1a). Genetic analysis of the suppressor revealed the presence of an insertion sequence element (IS2) within *yedV*, a gene coding for the histidine kinase of the uncharacterized YedV/YedW two-component system<sup>4</sup>. In close vicinity were two genes, *yedY* and *yedZ*, encoding, respectively, a periplasmic molybdopterin-containing oxidoreductase

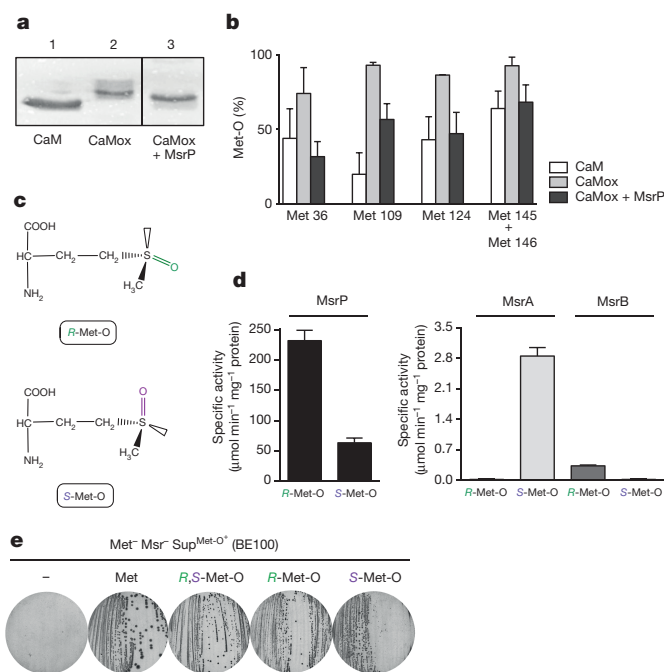
and its putative membrane redox partner<sup>5,6</sup>. YedY had been shown to reduce a variety of substrates *in vitro*, including trimethylamine N-oxide, and dimethyl, methionine and tetramethylene sulfoxides<sup>5</sup>. However, its physiological function had remained elusive, although a recent study in *Azospira suillum* suggested the homologous protein to



**Figure 1 | The MsrPQ system reduces free Met-O and is induced by HOCl.** **a**, JB590, a methionine auxotroph (Met<sup>-</sup>, numbered '1'), lacking all cytoplasmic Msrs (Met<sup>-</sup> Msr<sup>-</sup>, '2'), cannot grow on Met-O as the only Met source in contrast to suppressor BE100 (Met<sup>-</sup> Msr<sup>-</sup> Sup<sup>Met-O</sup>, '3'). Deletion of *yedY* (renamed *msrP*, '7'), *moeA* ('6') and *tatC* ('5'), but not of *trxA* ('4'), prevents the growth of BE100. **b**, The *yedYZ* operon is upregulated in BE100. The increase is observed both at the mRNA (quantitative PCR with reverse transcription (RT-qPCR), top; error bars, mean ± s.e.m.; *n* = 3) and protein (western blot, bottom) levels. The RT-qPCR primers were designed to quantify the *yedY-yedZ* mRNA. **c**, Immunoblot analysis showing that HOCl (2 mM), but not H<sub>2</sub>O<sub>2</sub> (1 mM), induces YedY synthesis in a wild-type strain. Images in **a–c** are representative of experiments made in biological triplicate. Uncropped blots are in Supplementary Fig. 1.

<sup>1</sup>WELBIO, Avenue Hippocrate 75, 1200 Brussels, Belgium. <sup>2</sup>de Duve Institute, Université catholique de Louvain, Avenue Hippocrate 75, 1200 Brussels, Belgium. <sup>3</sup>Brussels Center for Redox Biology, Avenue Hippocrate 75, 1200 Brussels, Belgium. <sup>4</sup>Aix-Marseille Université, CNRS, Laboratoire de Chimie Bactérienne, UMR 7283, Institut de Microbiologie de la Méditerranée, 31 Chemin Joseph Aiguier, 13009 Marseille, France. <sup>5</sup>Aix-Marseille Université, Centrale Marseille, CNRS, iSm2 UMR 7313, 13397, Marseille, France. <sup>†</sup>Present address: Department of Physics, Princeton University, Princeton, New Jersey 08544, USA.

\*These authors contributed equally to this work.



**Figure 2 | MsrP non-stereospecifically reduces protein-bound Met-O.**

**a**, Oxidation of Met in calmodulin (CaM) by  $\text{H}_2\text{O}_2$  leads to a mobility shift of the oxidized protein (CaMox); compare lane 2 with lane 1. Incubation of CaMox with MsrP and a reducing system involving dithionite and benzyl viologen restores the mobility (lane 3). **b**, MsrP can reduce Met-O in CaMox. The oxidation state of peptides containing either Met36, Met109, Met124 or Met145–146 was determined by LC–MS/MS. Error bars, mean  $\pm$  s.e.m.;  $n = 4$  for Met36, 109 and 145–146;  $n = 5$  for Met124. Met-O residues were detected in the untreated and MsrP-treated samples owing to limitations inherent to the methodology applied and oxidation of the samples during analytical handling. **c**, Representation of the two diastereoisomers of Met-O, *R*-Met-O and *S*-Met-O. **d**, MsrP exhibits activity towards both diastereoisomers (left), contrary to the stereospecific enzymes MsrA and MsrB (right). Specific activities were assayed using 64 mM of either *R*- or *S*-Met-O. Error bars, mean  $\pm$  s.d.;  $n = 3$ . **e**, The suppressor BE100 is able to grow on both isoforms of Met-O. Images in **a** and **e** are representative of experiments made in biological triplicate. The uncropped gel is in Supplementary Fig. 2.

be important for hypochlorous acid (HOCl) resistance<sup>7</sup>. We found that insertion of the IS2 led to a 100-fold increase in the levels of the *yedYZ* messenger RNA (mRNA) in strain BE100 and to higher YedY protein levels (Fig. 1b). Deletion of either *yedY* or *yedZ* prevented BE100 from growing on Met-O (Fig. 1a and Extended Data Table 1), while the simultaneous overproduction of YedY and YedZ, but not of YedY or YedZ alone, rendered the parental strain JB590 able to use Met-O (Extended Data Table 1). Altogether, these results indicated that the ability of the suppressor strain BE100 to reduce Met-O resulted from the increased synthesis of YedY and YedZ, implying that these two proteins function together as an Msr system. Growth of the BE100 strain was dependent on *moeA*, a gene required for the synthesis of molybdopterin cofactors, and on *tatC*, encoding a protein required for the translocation of metalloenzymes across the inner membrane (Fig. 1a). Exposure of wild-type cells to HOCl, but not to  $\text{H}_2\text{O}_2$ , induced the synthesis of YedY to levels comparable to those observed in BE100 (Fig. 1c), indicating that these proteins are specifically expressed in response to bleach stress. Interestingly, induction by HOCl was dependent on the presence of a functional YedV/YedW system (Extended Data Fig. 1).

All previously identified Msrs rely on electrons derived from NADPH via the thioredoxin (Trx) system for activity<sup>1</sup>. This was not the case for YedYZ, as deletion of *trxA*, encoding the Trx responsible for Msr recycling<sup>8</sup>, had no effect on the ability of BE100 to reduce

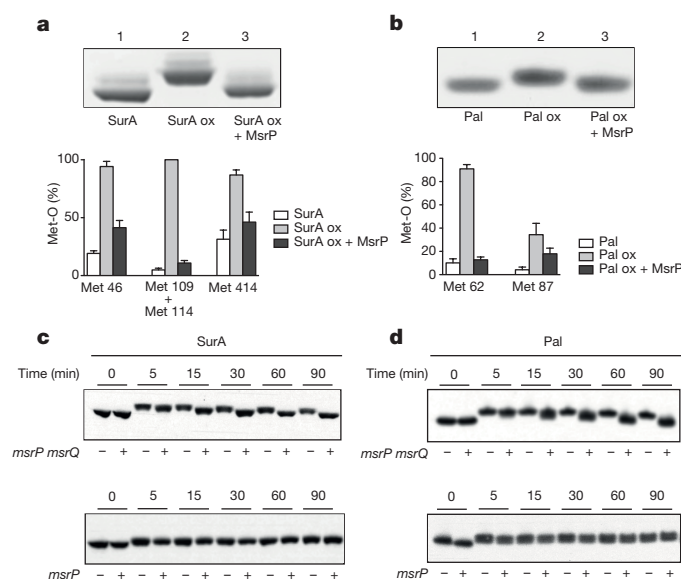
Met-O (Fig. 1a). As YedZ contains a *b*-type haem<sup>9</sup>, a cofactor typically associated with the quinone-oxidizing cytochrome *b* of the respiratory chain complexes, we considered the respiratory chain as a potential electron source. Deletion of *menA* and *ubiE*, two genes required for quinone synthesis, prevented BE100 from using Met-O (Extended Data Table 1), supporting a model in which YedZ uses electrons derived from the inner membrane pool of mature quinones to provide reducing equivalents to YedY (Extended Data Fig. 2). From now on, YedY and YedZ will be referred to as MsrP (for periplasm) and MsrQ (for quinone), respectively.

We then tested whether MsrPQ, in addition to free Met-O, also rescued Met-O residues present in proteins. Purified MsrP was shown to reduce *N*-acetyl-Met-O, a substrate mimicking protein-bound Met-O (Extended Data Fig. 3a), with a Michaelis constant ( $K_m$ ) of  $3.8 \pm 1.2$  mM, in line with the values reported for other Msrs<sup>10</sup>. Note that in the experiments involving purified MsrP, electrons were provided to the oxidoreductase by an inorganic system reducing molybdoenzymes<sup>11</sup>. Next, we tested the ability of MsrP to reduce oxidized calmodulin (CaMox), a substrate commonly used to assess Msr activity. We used a gel shift assay based on the reduced mobility exhibited in SDS–polyacrylamide gel electrophoresis (SDS–PAGE) by proteins containing Met-O<sup>12</sup>. Incubation of CaMox with MsrP restored its mobility, suggesting that MsrP was able to reduce Met-O residues in CaMox (Fig. 2a). This was confirmed by showing with liquid chromatography–tandem mass spectrometry (LC–MS/MS) that the oxidized Met residues that could be detected in CaMox were reduced back to levels similar to those observed in CaM after incubation with MsrP (Fig. 2b). Altogether, these results indicated that MsrP is able to reduce protein-bound Met-O.

Upon oxidation, two diastereoisomers of Met-O can form, referred to as *R* and *S*, owing to the asymmetric position of the oxidized sulfur atom in the lateral chain (Fig. 2c). All Msrs described so far exhibit stereospecificity, specifically reducing either the *R* (MsrB, MsrC) or the *S* isoform (MsrA, BisC). Using highly pure diastereoisomers (Extended Data Fig. 4), we found MsrP to exhibit activity towards both (Fig. 2d), with  $K_m$  values of  $25.7 \pm 4.7$  mM and  $8.0 \pm 2.7$  mM for *R*- and *S*-Met-O, respectively (Extended Data Fig. 3b). Accordingly, the BE100 suppressor strain was able to use *R*- and *S*-Met-O (Fig. 2e), in contrast to strains expressing single stereospecific Msrs (Extended Data Fig. 3c). Thus, MsrP is a new type of Msr with no stereospecificity.

To search for the physiological substrates of MsrP, periplasmic proteins from a  $\Delta msrP$  mutant were oxidized with HOCl, incubated with MsrP and subjected to a semi-quantitative two-dimensional LC–MS/MS analysis. Twenty proteins that had one or more HOCl-oxidized Met residues that MsrP could reduce were identified (Extended Data Table 2). Using gel shift assays in combination with LC–MS/MS analysis, we confirmed the ability of MsrP to reduce the chaperone SurA and the lipoprotein Pal (Fig. 3a, b). Altogether, these results established that MsrP is able to repair a wide panel of structurally and functionally diverse periplasmic proteins *in vitro*.

SurA is the primary periplasmic chaperone, escorting most  $\beta$ -barrel proteins to the outer membrane<sup>13,14</sup>. As HOCl-oxidized SurA loses its chaperone activity (Fig. 4a), we used this property to probe the physiological importance of the MsrPQ system. First, we showed that SurA could be oxidized *in vivo* by HOCl and that expression of the MsrPQ system, but not of MsrP alone, restored its mobility (Fig. 3c). Similar results were obtained for Pal (Fig. 3d), confirming that MsrP and MsrQ collaborate in the protection of SurA and Pal from oxidative damage. We then tested if the repair of SurA by MsrP, which restores the activity of the chaperone *in vitro* (Fig. 4a), was important to keep SurA active under HOCl stress. For this, we used a mutant strain lacking the chaperone Skp, in which SurA becomes essential<sup>15,16</sup>. We found that deleting *msrP* rendered the  $\Delta skp$  strain hypersensitive to HOCl (Fig. 4b), suggesting that oxidized, inactive SurA accumulates in the absence of MsrP. In agreement with this, the sensitivity of the  $\Delta skp \Delta msrP$

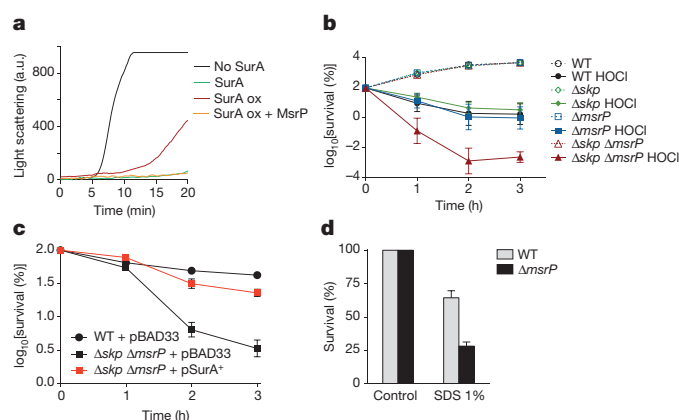


**Figure 3 | The MsrPQ system rescues oxidized Met residues in SurA and Pal.** **a, b**, Oxidation of SurA (SurA ox) and Pal (Pal ox) by  $H_2O_2$  leads to a mobility shift resulting from Met-O formation. Incubation with MsrP and the inorganic reducing system restores their mobility (top). The percentages of Met-O in the various samples were determined by LC-MS/MS analysis, confirming that MsrP reduces Met-O in SurA and Pal (bottom). Error bars, mean  $\pm$  s.e.m.;  $n = 3$ . Met-O residues were detected in the untreated and MsrP-treated samples owing to limitations inherent to the methodology applied and oxidation of the samples during analytical handling. **c, d**,  $\Delta msrPQ$  cells carrying *msrP* either alone or with *msrQ* under an isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG)-inducible promoter on a plasmid (pAG192 and pAG195, respectively) were grown with IPTG (100  $\mu$ M). Cells were treated with chloramphenicol (300  $\mu$ g  $ml^{-1}$ ) at an absorbance ( $A_{600nm}$ ) of 0.5 to block new protein synthesis and HOCl (3.5 mM) was added. Synthesis of MsrP and MsrQ together (top), but not of MsrP alone (bottom), restores SurA and Pal mobility. Images in **a–d** are representative of experiments made in biological triplicate. The small shift exhibited by SurA over time in the absence of MsrPQ could be due to a residual Msr activity, possibly an NADPH-dependent membrane-bound Msr activity previously detected<sup>21</sup>. Uncropped gels and blots are in Supplementary Fig. 3.

mutant to HOCl was suppressed by overexpression of SurA (Fig. 4c). Further highlighting the need to protect Met residues in periplasmic proteins, HOCl-pretreated  $\Delta msrP$  mutants were found to be more sensitive to SDS, a phenotype indicative of defects in the outer membrane (Fig. 4d)<sup>17</sup>.

The conservation of MsrPQ throughout Gram-negative bacteria (Extended Data Figs 5 and 6) illustrates the importance of having a Met-O reducing system in the periplasm. *Neisseria* species stand out as an exception in lacking MsrPQ. However, in these bacteria, evolutionary tinkering generated an envelope hybrid protein combining two classic stereospecific Msr domains<sup>18</sup>. A remarkable feature of MsrPQ is that its rescue activity depends on electrons provided by the respiratory chain. This represents an entirely novel way to provide reducing power for protein quality control in the envelope. Indeed, known reducing systems functioning in the periplasm use electrons provided by the inner membrane protein DsbD and Trx<sup>19</sup>. Hence, diverting electrons from the respiratory chain to control extracytosolic protein quality is an unprecedented link between metabolism and cellular integrity.

The chaperone SurA is one of the targets of the MsrPQ system. Having a protein folding helper under the control of a repair system reveals an additional layer in the complex control network of periplasmic protein quality. Testing if this system is an attractive target for antimicrobial development, as suggested by the colonization defect exhibited by the *msrP* mutant in *Campylobacter jejuni*<sup>20</sup>, will be the



**Figure 4 | The reducing activity of the MsrPQ system is important for envelope integrity.** **a**, Repair of oxidized SurA (SurA ox) by MsrP (SurA ox + MsrP) restores the ability of SurA to protect thermally unfolded citrate synthase from aggregation. The graph is representative of experiments made in biological triplicate (a.u., arbitrary units). **b**, While the wild-type (WT), the  $\Delta skp$  and the  $\Delta msrP$  strains are only moderately affected by exposure to HOCl (2 mM), the viability of the  $\Delta skp \Delta msrP$  mutant (in which SurA is essential) is decreased. Error bars, mean  $\pm$  s.e.m.;  $n = 3$ . **c**, The sensitivity of the  $\Delta skp \Delta msrP$  mutant to HOCl is suppressed by SurA overexpression. Error bars, mean  $\pm$  s.e.m.;  $n = 4$ . **d**, Pre-treatment with HOCl renders the  $\Delta msrP$  mutant hypersensitive to SDS, indicative of envelope defects. Error bars, mean  $\pm$  s.e.m.;  $n = 3$ .

subject of future research. By highlighting the importance of protecting proteins targeted to oxidizing compartments, our work calls for a detailed investigation of the process of Met-O reduction in the endoplasmic reticulum, where only an R-Met-O-specific MsrB has been identified<sup>3</sup>. As has long been speculated, a possibility would be that the endoplasmic reticulum contains an epimerase catalysing the interconversion of R- and S-Met-O. Alternatively, in light of the present study, the endoplasmic reticulum could contain a novel Met-O reducing system yet to be discovered.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 March; accepted 30 September 2015.

Published online 7 December 2015.

- Boschi-Muller, S., Gand, A. & Branlant, G. The methionine sulfoxide reductases: catalysis and substrate specificities. *Arch. Biochem. Biophys.* **474**, 266–273 (2008).
- Ezraty, B., Aussel, L. & Barras, F. Methionine sulfoxide reductases in prokaryotes. *Biochim. Biophys. Acta* **1703**, 221–229 (2005).
- Lee, B. C. & Gladyshev, V. N. The biological significance of methionine sulfoxide stereochemistry. *Free Radic. Biol. Med.* **50**, 221–227 (2011).
- Urano, H., Umezawa, Y., Yamamoto, K., Ishihama, A. & Ogasawara, H. Cooperative regulation of the common target genes between hydrogen peroxide-response YedVW and copper-response CusSR in *Escherichia coli*. *Microbiology* **161**, 729–738 (2015).
- Loschi, L. *et al.* Structural and biochemical identification of a novel bacterial oxidoreductase. *J. Biol. Chem.* **279**, 50391–50400 (2004).
- Workun, G. J., Moquin, K., Rothery, R. A. & Weiner, J. H. Evolutionary persistence of the molybdopyranopterin-containing sulfite oxidase protein fold. *Microbiol. Mol. Biol. Rev.* **72**, 228–248 (2008).
- Melnyk, R. A. *et al.* Novel mechanism for scavenging of hypochlorite involving a periplasmic methionine-rich peptide and methionine sulfoxide reductase. *MBio* **6**, e00233–15 (2015).
- Stewart, E. J., Aslund, F. & Beckwith, J. Disulfide bond formation in the *Escherichia coli* cytoplasm: an *in vivo* role reversal for the thioredoxins. *EMBO J.* **17**, 5543–5550 (1998).
- Brox, S. J., Rothery, R. A., Zhang, G., Ng, D. P. & Weiner, J. H. Characterization of an *Escherichia coli* sulfite oxidase homologue reveals the role of a conserved active site cysteine in assembly and function. *Biochemistry* **44**, 10339–10348 (2005).
- Tarrago, L. & Gladyshev, V. N. Recharging oxidative protein repair: catalysis by methionine sulfoxide reductases towards their amino acid, protein, and model substrates. *Biokhimiia* **77**, 1097–1107 (2012).



11. Lowe, R. H. & Evans, H. J. Preparation and some properties of a soluble nitrate reductase from *Rhizobium japonicum*. *Biochim. Biophys. Acta* **85**, 377–389 (1964).
12. Le, D. T. *et al.* Analysis of methionine/selenomethionine oxidation and methionine sulfoxide reductase function using methionine-rich proteins and antibodies against their oxidized forms. *Biochemistry* **47**, 6685–6694 (2008).
13. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* **2**, a000414 (2010).
14. Goemans, C., Denoncin, K. & Collet, J. F. Folding mechanisms of periplasmic proteins. *Biochim. Biophys. Acta* **1843**, 1517–1528 (2014).
15. Sklar, J. G., Wu, T., Kahne, D. & Silhavy, T. J. Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in *Escherichia coli*. *Genes Dev.* **21**, 2473–2484 (2007).
16. Denoncin, K., Schwalm, J., Vertommen, D., Silhavy, T. J. & Collet, J. F. Dissecting the *Escherichia coli* periplasmic chaperone network using differential proteomics. *Proteomics* **12**, 1391–1401 (2012).
17. Ruiz, N., Falcone, B., Kahne, D. & Silhavy, T. J. Chemical conditionality: a genetic strategy to probe organelle assembly. *Cell* **121**, 307–317 (2005).
18. Brot, N. *et al.* The thioredoxin domain of *Neisseria gonorrhoeae* PilB can use electrons from DsbD to reduce downstream methionine sulfoxide reductases. *J. Biol. Chem.* **281**, 32668–32675 (2006).
19. Cho, S. H. & Collet, J. F. Many roles of the bacterial envelope reducing pathways. *Antioxid. Redox Signal.* **18**, 1690–1698 (2013).
20. Hitchcock, A. *et al.* Roles of the twin-arginine translocase and associated chaperones in the biogenesis of the electron transport chains of the human pathogen *Campylobacter jejuni*. *Microbiology* **156**, 2994–3010 (2010).
21. Spector, D., Etienne, F., Brot, N. & Weissbach, H. New membrane-associated and soluble peptide methionine sulfoxide reductases in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* **302**, 284–289 (2003).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank A. Boujdat, G. Herinckx and J.-P. Szikora for technical and computational help and the members of the Barras and Collet laboratories for discussions. We are indebted to M. Sabaty and D. Pignol for sharing unpublished information, to T. Silhavy, T. Palmer, E. Bouveret and D. Hughes for providing strains and plasmids, to T. Lowther and M. Réglier for advice and discussions and to N. Typas, T. Mignot, J. Messens, J. Bardwell and F.-A. Wollman for reading the manuscript and providing comments. A.G. and J.S. are research fellows of the Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture, P.L. is 'Chargée de Recherche' and J.-F.C. is 'Maître de Recherche' of the Fonds de la Recherche Scientifique-FNRS (FRS-FNRS). E.O. is supported by a grant from the Indo-French Center for the Promotion of Advanced Research CEFIPRA '(5105-2)'. R.A. is supported by the Fonds Maurange, Fondation Roi Baudouin. This work was supported, in part, by grants from the FRS-FNRS and from the European Research Council (FP7/2007–2013) ERC independent researcher starting grant 282335–Sulfenic to J.-F.C. and funding by the Centre National de la Recherche Scientifique (CNRS), Fondation pour la Recherche Médicale (FRM) and Aix-Marseille Université to the F.B. team.

**Author Contributions** F.B., J.-F.C., A.G. and B.E. wrote the paper. A.G., B.E., C.H., A.V., L.E., J.-F.C. and F.B. designed and performed the experiments. A.G., B.E., C.H., J.B., P.L. and J.S. constructed the strains and cloned the constructs. F.B., J.-F.C., A.G., B.E. and C.H. analysed and interpreted the data. D.V. performed MS analyses. E.O. and O.I. prepared the diastereoisomers. R.A. performed bioinformatic analyses. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.B. ([barras@imm.cnrs.fr](mailto:barras@imm.cnrs.fr)) or J.-F.C. ([jfc@uclouvain.be](mailto:jfc@uclouvain.be)).

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**Strains and microbial techniques.** The strains used in this study are listed in Supplementary Table 1. Unless otherwise specified, for all deletion mutants, the corresponding alleles from the Keio collection<sup>22</sup> were transferred into the MC4100 wild-type strain using P1 transduction standard procedures<sup>23</sup> and checked by PCR. To excise the resistance cassette, we used pCP20 (refs 22, 24). Strain AG227, deleted for the entire *yedYZ* operon, was constructed as follows. First, a *cat-sacB* cassette, encoding chloramphenicol acetyl transferase and *SacB*, a protein conferring sensitivity to sucrose, was amplified from strain CH1990 using primers *yedYZ::cat-sacB\_Fw* and *yedYZ::cat-sacB\_Rv*. The resulting PCR product shared a 40-base-pair (bp) homology to the 5' untranslated region of *yedY* (*msrP*) and to the 3' untranslated region of *yedZ* (*msrQ*) at its 5' and 3' ends, respectively. After purification, the PCR product was transformed by electroporation into CH1940. These cells harbour the pSIM5-tet vector, which encodes the Red recombination system proteins Gam, Beta and Exo under the control of the temperature-sensitive repressor cI859, encoded by the same vector. Induction of the Gam, Beta and Exo proteins was induced by shifting the cells to 42 °C for 15 min before making them electrocompetent. Recombinant cells were selected on chloramphenicol-containing plates (25 µg ml<sup>-1</sup>) at 37 °C for 16 h. At this temperature, the pSIM5-tet vector, which has a temperature-sensitive origin of replication, is lost. Colonies were also tested for the presence of the *cat-sacB* cassette by negative selection on sucrose-containing media (5% sucrose, no NaCl). Finally, we verified that the *cat-sacB* cassette replaced the *msrPQ* operon in the resulting strain (AG219) by sequencing across the junctions. The *cat-sacB* cassette was subsequently moved from AG219 to TP1004 by P1 transduction. The *cat-sacB* cassette was eliminated from the resulting strain (AG220) by transforming it with the pSIM5-tet plasmid, electroporating it with the oligonucleotide Delta\_*yedYZ* (300 ng) and performing lambda red recombination as described above. Recombinants were selected on sucrose-containing media at 30 °C for 16 h. To eliminate the plasmid, the selected colonies were grown at 37 °C for 16 h. Loss of the cassette in the resulting AG227 strain was verified by positive (sucrose resistance) and negative (chloramphenicol sensitivity) selection and by PCR.

The *msrQ* deletion mutant (strain BE105) was generated using the PCR knock-out method developed in ref. 24. Briefly, a DNA fragment containing the *cat* gene flanked with the homologous sequences found upstream and downstream of the *yedZ* gene was PCR-amplified using pKD3 as template and the oligonucleotides P1\_Up\_YedZ and P2\_Down\_YedZ. Strain BE100, carrying plasmid pKD46, was then transformed by electroporation with the amplified linear fragment. Chloramphenicol-resistant clones were selected and verified by PCR.

The *msrP::lacZ* fusion was constructed using the method described in ref. 25. Briefly, the *msrP* promoter region lying between nucleotide -797 and nucleotide +63, using the A nucleotide within the initiation triplet as a reference, was amplified by PCR with the appropriate oligonucleotides (*lacI-msrP*<sub>forward</sub> and *lacZ-msrP'*<sub>reverse</sub>). Using mini-lambda-mediated recombinering, the PCR product was then directly recombined with the chromosome of a modified *E. coli* wild-type strain (PM1205), carrying a P<sub>BAD</sub>-*cat-sacB* cassette inserted in front of *lacZ*, at the ninth codon. Recombinants were selected for loss of the *cat-sacB* genes, resulting in the translational fusion of *msrP* to *lacZ*.

**Plasmid construction.** The plasmids and primers used in this study are listed in Supplementary Tables 2 and 3, respectively. The YedY-His<sub>6</sub> (MsrP-His<sub>6</sub>) expression vector was constructed as follows. Site-directed mutagenesis using primers pTAC\_NdeI\_Fw and pTAC\_NdeI\_Rv was performed using pTAC-MAT-Tag-2 as template to introduce an NdeI restriction site in the vector, yielding vector pAG177. *yedY* (*msrP*) DNA was amplified from the chromosome (MC4100) using primers pTAC\_*yedY*\_Fw and pTAC\_*yedY*-His<sub>6</sub>\_Rv, which resulted in the fusion of a His<sub>6</sub> tag coding sequence at the 3' end. The PCR product was subsequently cloned into pAG177 using NdeI and BglII restriction sites, generating plasmid pAG178. To construct IPTG-inducible pTAC-MAT-Tag-2 vectors expressing either MsrP (without tag) or both MsrP and MsrQ, we first amplified the corresponding coding DNA sequences (*msrP* or the *msrPQ* operon) from the chromosome of strain MC4100 using primer pairs pTAC\_*yedY*\_Fw/ pTAC\_*yedY*\_Rv and pTAC\_*yedY*\_Fw/ pTAC\_*yedZ*\_Rv, respectively. The PCR products were then cloned into pAG177 using restriction sites NdeI and BglII, yielding pAG192 (MsrP) and pAG195 (MsrPQ). The complementation pAM238 vectors constitutively expressing either MsrP or MsrQ alone (without tag) or both MsrP and MsrQ were constructed as follows. We first amplified the corresponding coding DNA sequences (*msrP*, *msrQ* or the *msrPQ* locus) in addition to a 50 bp upstream region from each start codon (to include a ribosomal binding site) from the chromosome of strain MG1655 using primer pairs pAM238\_*yedY*\_Fw/ pAM238\_*yedY*\_Rv, pAM238\_*yedZ*\_Fw/ pAM238\_*yedZ*\_Rv and pAM238\_*yedY*\_Fw/ pAM238\_*yedZ*\_Rv, respectively.

The PCR products were then cloned into pAM238 using restriction sites KpnI and PstI, yielding pAG264 (MsrP), pAG275 (MsrQ), and pAG265 (MsrPQ).

The vector allowing the arabinose-inducible expression of *SurA* was constructed as follows. The *surA*-encoding DNA and its 50 bp upstream region (to include a ribosomal binding site) were amplified from the chromosome of strain MG1655 using the primer pair *surA\_Fw/surA\_Rv*. The PCR product was then cloned into pBAD33 using restriction sites KpnI and XbaI, yielding vector pAG290.

**Analysis of the *yedYZ* operon expression by RT-qPCR.** Expression levels of the *yedYZ* (*msrPQ*) mRNA were assessed in M63 minimal medium supplemented with 0.5% glycerol, 0.15% casamino acids, 1 mM MgSO<sub>4</sub>, 1 mM MoNa<sub>2</sub>O<sub>4</sub>, 17 µM Fe<sub>2</sub>(SO<sub>4</sub>)<sub>3</sub> and vitamins (thiamine 10 µg ml<sup>-1</sup>, biotin 1 µg ml<sup>-1</sup>, riboflavin 10 µg ml<sup>-1</sup> and nicotinamide 10 µg ml<sup>-1</sup>). Overnight cultures of MG1655 were diluted to A<sub>600 nm</sub> = 0.04 in fresh M63 minimal medium (100 ml) and cultured aerobically at 37 °C until A<sub>600 nm</sub> = 0.8. Cells (10 ml) were then pelleted, resuspended in TriPure (Roche) and homogenized. After mixing with chloroform, RNA was isolated by centrifugation (15 min, 15,700g, 4 °C), precipitated with isopropanol, washed with ethanol 70%, dried and finally resuspended in DEPC water. Any residual DNA was eliminated by treatment of the sample with DNase (Turbo DNA-free Kit, Ambion). A RevertAid RT kit (Thermo Scientific) was used to generate complementary DNA (cDNA) from 1 µg RNA extracted from each of the cultured strains. cDNAs were then diluted 1/10 and submitted to qPCR, using a qPCR Core kit for SYBR Green I No ROX (Eurogentec) and a MyiQ Single-Colour Real-Time PCR Detection System (Bio-Rad). Expression levels of *yedYZ* were normalized to the expression of *gapA*. Primers used for qPCR analysis were qPCR\_*yedYZ*\_Fw and qPCR\_*yedYZ*\_Rv for *yedYZ*, and qPCR\_*gapA*\_Fw and qPCR\_*gapA*\_Rv for *gapA* (Supplementary Table 3).

**Immunoblot analysis of MsrP expression.** Synthesis of MsrP in strains JB590 and BE100 was assessed as follows. Overnight cultures were diluted to A<sub>600 nm</sub> = 0.04 in fresh M63 minimal medium (100 ml) and cultured aerobically at 37 °C until A<sub>600 nm</sub> = 0.8. Nine hundred microlitres of each culture were then precipitated with 10% ice-cold trichloroacetic acid (TCA), pellets were washed with ice-cold acetone, dried, resuspended and heated at 95 °C in Laemmli SDS sample buffer (SB buffer) (2% SDS, 10% glycerol, 60 mM Tris-HCl, pH 7.4, 0.01% bromophenol blue), and loaded on an SDS-PAGE gel for immunoblot analysis. The protein amounts loaded were standardized by taking into account the A<sub>600 nm</sub> values of the cultures.

To monitor the MsrP expression levels after NaOCl or H<sub>2</sub>O<sub>2</sub> treatment, overnight cultures of wild-type cells (MG1655) were diluted to A<sub>600 nm</sub> = 0.04 in fresh lysogeny broth (LB) medium (100 ml) and grown aerobically at 37 °C to A<sub>600 nm</sub> = 0.5. NaOCl (2 mM) or H<sub>2</sub>O<sub>2</sub> (1 mM) was then added to the cultures. Samples were TCA-precipitated, washed with ice-cold acetone, dried, suspended in SB buffer, heated at 95 °C and loaded on an SDS-PAGE gel for immunoblot analysis. The protein amounts loaded were standardized by taking into account the A<sub>600 nm</sub> values of the cultures. The specificity of the anti-MsrP antibody was verified (Supplementary Fig. 5).

**Preparation of pure diastereoisomeric forms of Met-O.** L-Methionine sulfoxide ([α]<sub>D</sub><sup>24</sup> = +14.3° (water)), triethylamine (>99%) and methanol (>99.6%) were obtained from Sigma-Aldrich, picric acid from Prolabo and D<sub>2</sub>O from SDS. Water was purified using Millipore Elix Essential 3 apparatus. <sup>1</sup>H and <sup>13</sup>C NMR were recorded on a Bruker Avance III Nanobay spectrometer (<sup>1</sup>H: 400 MHz; <sup>1</sup>H/<sup>13</sup>C: 100 MHz). Chemical shifts (δ) were referenced to dioxane (<sup>1</sup>H: δ = 3.75 p.p.m.; <sup>13</sup>C: δ = 67.19 p.p.m.)<sup>26</sup>, which was added as an internal reference; resonances are detailed as follows: <sup>1</sup>H, δ in parts per million (multiplicity, J-coupling in hertz, integration, signal attribution); <sup>1</sup>H/<sup>13</sup>C, δ in parts per million (signal attribution). For each diastereoisomer, chemical shifts are similar to those previously reported<sup>27</sup>. <sup>13</sup>C resonance assignments were confirmed by heteronuclear single quantum coherence experiments. Optical rotations were measured on an Anton Paar Modular Circular Polarimeter 200 instrument at 25 °C and 589 nm from aqueous solution containing 0.8–1.2 g per 100 ml of L-methionine sulfoxide. The values reported are the average and s.d. relative to three independent measurements recorded on distinct solutions.

The commercial mixture of diastereoisomers was separated following the previously reported method<sup>28</sup>. Briefly, 10 ml of water was added to L-methionine sulfoxide (1.333 g, 8.069 mmol) and picric acid (1.849 g, 8.071 mmol). The suspension was heated to reflux until complete dissolution and then slowly cooled to room temperature (~25 °C). The suspension was filtered on a sintered funnel and the solid was washed with cold water (10 ml in total). Both the solid (*dextro*) and filtrate (*levo*) were collected separately for further purification.

*Dextro*. To the dried solid, 20 ml of water were added and the mixture was heated to reflux then allowed to cool slowly to room temperature. The solid was filtered out, washed with 10 ml water and dried. Again, 11 ml of methanol were added to the resulting solid and the mixture heated to reflux. After slow cooling, the yellow crystals were filtered, washed with 5 ml methanol and dried. A portion was used for structure determination by X-ray analysis. To the *dextrogyre* picrate

salt (1.345 g, 3.42 mmol), ~1.1 equivalents of triethylamine were added as a dilute aqueous solution (22 ml, 175 mM, 3.85 mmol). Subsequently, 200 ml of acetone were added portion-wise to the above stirring suspension and a white solid precipitated. This was filtered, washed, triturated with acetone and finally dried in vacuum (533 mg, 80%).

**Levo.** The volume of the filtrate was reduced in vacuum at 40 °C to about 3–4 ml to obtain a saturated solution and a small amount of precipitate. Then, 1.5 ml of water were added, the suspension was filtered and the solid washed with minimal water (2 ml). The whole step was repeated once (reduce the volume, dilute, filter and wash), and the resulting solution was then completely dried in vacuum. To the resulting yellow residue, 15 ml of methanol were added and the suspension was heated to reflux. In our hands, no solid precipitated upon cooling (in contrast with the reported method<sup>28</sup>); therefore the solution was dried again in vacuum. Following the same protocol as before, to the levogyre-enriched picrate salt (1.354 g, 3.44 mmol), ~1.1 equivalents of triethylamine were added as a concentrated aqueous solution (3.8 ml, 1 M, 3.8 mmol). Afterwards, 200 ml of acetone were added portion-wise and a white solid precipitated. This was filtered, washed, triturated with acetone and finally dried in vacuum (515 mg, 77%).

**Dextro (L-methionine-S-sulfoxide):**  $[\alpha]_D^{25} = +99.2 \pm 1.5^\circ$  (water);  $^1\text{H}$  NMR (400 MHz,  $\text{D}_2\text{O}$   $pD = 6.5$ ): 3.88 (t,  $^3J = 6.3$ , 1H,  $\text{H}_{\alpha\text{S}}$ ), 3.02 (m, 2H,  $\text{H}_{\gamma\text{S}}$ ), 2.74 (s, 3H,  $\text{H}_{\text{E}\text{S}}$ ), 2.31 (dd,  $J = 14.4$ , 7.6, 2H,  $\text{H}_{\beta\text{S}}$ );  $\{^1\text{H}\}^{13}\text{C}$  NMR (100 MHz,  $\text{D}_2\text{O}$ ): 173.8 ( $\text{COO}_{\text{S}}$ ), 54.0 ( $\text{C}_{\alpha\text{S}}$ ), 48.9 ( $\text{C}_{\gamma\text{S}}$ ), 37.2 ( $\text{C}_{\text{E}\text{S}}$ ), 24.4 ( $\text{C}_{\beta\text{S}}$ ). Literature values from ref. 28:  $[\alpha]_D^{25} = +99^\circ$  (water), from ref. 27:  $[\alpha]_D = +98.2^\circ$  (water, room temperature);  $^1\text{H}$  NMR (300 MHz,  $\text{D}_2\text{O}$ ): 4.10 (m, 1H), 3.08–2.78 (m, 2H), 2.59 (s, 3H), 2.32–2.13 (m, 2H);  $^{13}\text{C}$  NMR (75 MHz,  $\text{D}_2\text{O}$ ): 171.1, 52.0, 48.3, 37.0, 23.5.

**Levo (L-methionine-R-sulfoxide):**  $[\alpha]_D^{25} = -72.7 \pm 0.5^\circ$  (water);  $^1\text{H}$  NMR (400 MHz,  $\text{D}_2\text{O}$   $pD = 6.5$ ): 3.86 (t,  $^3J = 6.3$ , 1H,  $\text{H}_{\alpha\text{R}}$ ), 3.12 (ddd,  $J = 13.4$ , 9.6, 7.0, 1H,  $\text{H}_{\gamma\text{R}}$  or  $\text{H}_{\gamma\text{R}2}$ ), 3.02 (m, 2H,  $\text{H}_{\gamma\text{S}}$ ), 2.93 (ddd,  $J = 13.5$ , 9.1, 6.8, 1H,  $\text{H}_{\gamma\text{R}}$  or  $\text{H}_{\gamma\text{R}2}$ ), 2.74 (s, 3H,  $\text{H}_{\text{E}\text{R}}$ ), 2.31 (m, 2H,  $\text{H}_{\beta\text{R}}$ );  $\{^1\text{H}\}^{13}\text{C}$  NMR (100 MHz,  $\text{D}_2\text{O}$ ): 173.9 ( $\text{COO}_{\text{R}}$ ), 54.2 ( $\text{C}_{\alpha\text{R}}$ ), 54.0 ( $\text{C}_{\alpha\text{S}}$ ), 48.9 ( $\text{C}_{\gamma\text{R}}$ ), 37.2 ( $\text{C}_{\text{E}\text{S}}$ ), 37.0 ( $\text{C}_{\text{E}\text{R}}$ ), 24.4 ( $\text{C}_{\beta\text{R}}$ ). Literature values from ref. 28:  $[\alpha]_D^{26} = -71.6^\circ$  (water), from ref. 27:  $[\alpha]_D = -78^\circ$  (water, room temperature);  $^1\text{H}$  NMR (300 MHz,  $\text{D}_2\text{O}$ ): 4.10 (m, 1H), 3.08–2.78 (m, 2H), 2.59 (s, 3H), 2.32–2.13 (m, 2H);  $^{13}\text{C}$  NMR (75 MHz,  $\text{D}_2\text{O}$ ): 171.1, 52.1, 48.4, 37.0, 23.7.

In the  $^1\text{H}$  NMR spectra, the resonance centred at 3.02 p.p.m. was attributed to the S-enantiomer. The relative integral values suggest that R-Met-O is contaminated by 3% of the S-diastereoisomer. Moreover, comparing the measured  $[\alpha]_D^{25}$  values with those reported in ref. 27, the data are consistent with the presence of 3% S-diastereoisomer as a contaminant. Such purity is in line with previous reports using the same separation method<sup>28,29</sup>. The absolute configuration of the L-methionine-S-sulfoxide was confirmed by X-ray structural analysis and matches previous assignments<sup>27,30</sup>.

**Synthesis of N-acetyl-Met-O.** To synthesize N-acetyl-Met-O, Met-O (30 mg; Sigma-Aldrich) was solubilized in 2 ml 100% acetic acid. After addition of 2 ml of 97% acetic anhydride, the resulting mixture was incubated 2 h at 23 °C. Then, 2 ml of water were added and the mixture was lyophilized overnight. Finally, the lyophilized N-acetyl-Met-O was washed three times with 6 ml of water, re-lyophilized and suspended in 500 mM  $\text{Na}_2\text{HPO}_4$ , pH 9.0 to a final concentration of 1.5 M. The pH was then adjusted to 7 with NaOH.

**Kinetic analysis of MsrP activity.** The MsrP reductase activity was followed spectrophotometrically at 600 nm by monitoring the substrate-dependent oxidation of reduced benzyl viologen, serving as an electron donor. Reactions were performed anaerobically at 30 °C in degassed and nitrogen-flushed 50 mM MOPS, pH 7.0 using stoppered cuvettes. Benzyl viologen was used at a final concentration of 0.4 mM (molar extinction coefficient,  $\epsilon$ , of reduced benzyl viologen =  $7,800 \text{ M}^{-1} \text{ cm}^{-1}$ ) and reduced with sodium dithionite. The final reaction volume was kept constant, with the ordered addition of benzyl viologen, sodium dithionite, 1–32 mM N-acetyl-methionine sulfoxide (NacMet-O) and 10 mM MsrP-His<sub>6</sub>. The concentrations used for the R- and S-Met-O diastereoisomers were 1–64 mM. The Michaelis–Menten parameters (maximum velocity ( $V_{\text{max}}$ ) and  $K_{\text{m}}$ ) were determined using Graphpad Prism software.

**Analysis of MsrA and MsrB activities.** The reductase activities of MsrA and MsrB were followed spectrophotometrically at 340 nm by monitoring the substrate-dependent oxidation of NADPH ( $\epsilon = 6,220 \text{ M}^{-1} \text{ cm}^{-1}$ ). Reactions were performed at 37 °C in HEPES–KOH 20 mM, pH 7.4, NaCl 10 mM, and the final reaction volumes were kept constant, with the ordered addition of 250  $\mu\text{M}$  NADPH (Roche), 2.6  $\mu\text{M}$  TrxR, 40  $\mu\text{M}$  Trx, 64 mM substrate and 1.5  $\mu\text{M}$  of either MsrA or MsrB.

**Identification of the periplasmic proteins repaired by MsrP using two-dimensional LC–MS/MS.** The identification of the MsrP substrates was performed as follows. AG89 cells (2L) were grown aerobically at 37 °C in terrific broth to  $A_{600 \text{ nm}} = 0.8$ . Periplasmic extracts were prepared as described previously<sup>31</sup>. Briefly, cells were pelleted by centrifugation at 3,000g for 20 min at 4 °C and incubated on

ice with gentle shaking for 30 min in 100 mM Tris–HCl, pH 8.0, 20% sucrose, 1 mM EDTA. This mixture also contained 20 mM N-ethylmaleimide to alkylate reduced cysteine residues in proteins to prevent their subsequent oxidation. Periplasmic proteins were then isolated by centrifugation of the cells at 3,000g for 20 min at 4 °C. The periplasmic extract was subsequently concentrated by ultrafiltration in an Amicon cell (3,000 Da cutoff, YM-3 membrane) and loaded on a PD-10 column (GE Healthcare) equilibrated with 50 mM NaPi, pH 8.0, 50 mM NaCl. After concentration using a 5 kDa cutoff Vivaspinn 4 (Sartorius) concentrator, the extract was finally separated in three samples. Two samples were incubated 10 min at 37 °C with 2 mM NaOCl whereas the third was left untreated to serve as reduced control. NaOCl was then removed by gel filtration using a NAP-5 column (GE Healthcare) equilibrated with 50 mM MOPS, pH 7.0. The untreated sample was also subjected to the NAP-5 gel filtration.

One of the NaOCl-oxidized fractions was then reduced *in vitro* by incubation for 1 h at 37 °C with 10  $\mu\text{M}$  MsrP, 10 mM benzyl viologen and an excess of sodium dithionite. The other NaOCl-oxidized fraction, used as an oxidized control, and the non-oxidized fraction were incubated with 10 mM benzyl viologen and an excess of sodium dithionite but without MsrP. The three samples were then de-salted by dialysis against 50 mM MOPS, pH 7.0 by using Slide-A-Lyzer 3,500 MWCO G2 cassettes (Thermo Scientific). The three samples (500  $\mu\text{g}$ ) were precipitated by adding TCA to a final concentration of 10% w/v. The resulting pellets were washed with ice-cold acetone, dried in a Speedvac, suspended in 0.1 M  $\text{NH}_4\text{HCO}_3$ , pH 8.0, digested overnight at 30 °C with 3  $\mu\text{g}$  sequencing-grade trypsin, and analysed by two-dimensional LC–MS/MS essentially as described<sup>32</sup>. Briefly, peptides were first separated on a first-dimension hydrophilic interaction liquid chromatography (HILIC) column with a reverse acetonitrile gradient and 25 fractions of 1 ml collected (2 min per fraction). After drying, peptides were analysed by LC–MS/MS on a C18 column. The MS scan routine was set to analyse by MS/MS the five most intense ions of each full MS scan; dynamic exclusion was enabled to assure detection of co-eluting peptides.

**Protein identification by mass spectrometry.** Raw data collection of approximately 230,000 MS/MS spectra per two-dimensional LC–MS/MS experiment was followed by protein identification using SEQUEST. All MS raw files have been deposited in the ProteomeXchange Consortium<sup>33</sup> via the PRIDE partner repository with the data set identifier PXD002804. In detail, peak lists were generated using extract-msn (ThermoScientific) within Proteome Discoverer 1.4.1. From raw files, MS/MS spectra were exported with the following settings: peptide mass range 350–5,000 Da; minimal total ion intensity 500. The resulting peak lists were searched using SequestHT against a target-decoy *E. coli* protein database (release 07.01.2008, 8,678 entries comprising forward and reverse sequences) obtained from Uniprot. The following parameters were used: trypsin was selected with proteolytic cleavage only after arginine and lysine, number of internal cleavage sites was set to 1, mass tolerance for precursors and fragment ions was 1.0 Da, and considered dynamic modifications were +15.99 Da for oxidized methionine and +125.12 Da for N-ethylmaleimide on cysteines. Peptide matches were filtered using the  $q$  value and posterior error probability calculated by the Percolator algorithm ensuring an estimated false positive rate below 5%. The filtered SEQUEST HT output files for each peptide were grouped according to the protein from which they were derived using the multiconsensus results tool within Proteome Discoverer. Then the values of the spectral matches of only Met-containing peptides were combined from the three two-dimensional LC–MS/MS experiments and exported in a Microsoft Excel spreadsheet, with the rows referring to the peptide sequences and the columns to the fractions of the HILIC column. Oxidation of Met residues to Met-O by NaOCl causes a hydrophilic shift, which influences their retention time and makes them elute later (4–8 min) than their reduced counterpart on a HILIC column. If these Met-O are reduced by MsrP, they will then show a hydrophobic shift and elute at the same retention time on the HILIC column as in the control sample. By comparing the retention times and the number of peptide spectral matches of the Met-O-containing peptides in a periplasmic extract under three experimental conditions (control, oxidized by NaOCl with and without MsrP), one can identify 'bona fide' potential MsrP substrates.

**Protein expression and purification.** TP1004 cells harbouring plasmid pAG178 and overexpressing MsrP-His<sub>6</sub> protein were grown aerobically at 30 °C in terrific broth (Sigma-Aldrich) supplemented with sodium molybdate (1.5 mM) and ampicillin (200  $\mu\text{g ml}^{-1}$ ). When cells reached  $A_{600 \text{ nm}} = 0.8$ , expression was induced with 0.1 mM IPTG for 3 h. Periplasmic proteins were then extracted as in ref. 32. MsrP-His<sub>6</sub> was then purified by loading the periplasmic extract on a 1 ml HisTrap FF column (GE Healthcare) equilibrated with buffer A (NaPi 50 mM, pH 8.0, NaCl 300 mM). After washing the column with buffer A, MsrP-His<sub>6</sub> was eluted by applying a linear gradient of imidazole (from 0 to 300 mM) in buffer A. The fractions containing MsrP-His<sub>6</sub> were pooled, concentrated using a 5 kDa cutoff Vivaspinn



15 (Sartorius) device and de-salted on a PD-10 column (GE Healthcare) equilibrated with 50 mM NaPi, pH 8.0, 150 mM NaCl.

VU1 CaM, MsrA and MsrB were expressed and purified as described previously<sup>34,35</sup>.

Trx was expressed and purified as follows. BL21 (DE3) cells harbouring plasmid pMD205, overexpressing Trx with a carboxy (C)-terminal His<sub>6</sub> tag, were grown aerobically at 37 °C in LB supplemented with kanamycin (50 µg ml<sup>-1</sup>). Expression was induced at  $A_{600\text{ nm}} = 0.6$  with 1 mM IPTG for 3 h. Cells were then pelleted, resuspended in buffer A (NaPi 50 mM, pH 8.0, NaCl 300 mM) and disrupted by two passes through a French pressure cell at 12,000 psi. The lysate was then centrifuged at 30,000g and at 4 °C for 45 min, to remove cell debris, and Trx was purified as described for MsrP-His<sub>6</sub>. Ni-NTA-purified Trx was then loaded on a 120 ml HiLoad 16/60 Superdex 75 PG column (GE Healthcare) previously equilibrated with HEPES-KOH 50 mM, pH 7.4, NaCl 100 mM. The resulting Trx-containing fractions were pooled and concentrated using a 5 kDa cutoff Vivaspinn 15 device.

Thioredoxin reductase (TrxR) was expressed and purified as follows. BL21 (DE3) cells harbouring plasmid pPL223-2, overexpressing TrxR with an amino (N)-terminal His<sub>6</sub> tag, were grown aerobically at 37 °C in LB supplemented with ampicillin (200 µg ml<sup>-1</sup>). Expression was induced at  $A_{600\text{ nm}} = 0.6$  with 1 mM IPTG for 3 h. Protein extraction was performed as described for Trx and purification was performed as described for MsrP-His<sub>6</sub>.

BL21 (DE3) cells harbouring plasmid pKD11, overexpressing SurA with a C-terminal His<sub>6</sub> tag, were grown aerobically at 37 °C in LB supplemented with kanamycin (50 µg ml<sup>-1</sup>). Expression was induced at  $A_{600\text{ nm}} = 0.6$  with 1 mM IPTG for 3 h. Protein extraction and purification were performed as described for MsrP-His<sub>6</sub>.

MG1655 cells harbouring plasmid pKD84, overexpressing SurA with a C-terminal Strep-tag, were grown aerobically at 37 °C in LB supplemented with ampicillin (200 µg ml<sup>-1</sup>). Expression was induced at  $A_{600\text{ nm}} = 0.7$  with a final concentration of 200 µg l<sup>-1</sup> anhydrotetracycline (AHT) for 5 h. Protein extraction was performed as described for MsrP-His<sub>6</sub>. SurA-Strep was then purified by loading the periplasmic extract on a 5 ml Strep-Tactin Superflow cartridge H-PR (IBA) equilibrated with buffer A (Tris-HCl 100 mM, pH 8.0, NaCl 150 mM, EDTA 1 mM). After washing the column with buffer A, SurA-Strep was eluted by applying a linear gradient of desthiobiotin (from 0 to 2.5 mM) in buffer A. The fractions containing SurA-Strep were pooled, concentrated using a 5 kDa cutoff Vivaspinn 15 (Sartorius) device and de-salted on a PD-10 column (GE Healthcare) equilibrated with 50 mM NaPi, pH 8.0, 150 mM NaCl.

A modified version of Pal lacking the signal sequence and in which the first cysteine of the lipobox was replaced by an alanine (Pal<sub>C1A</sub>) was expressed with an N-terminal His<sub>6</sub> tag from the pEB0513 vector in BL21 (DE3) cells. Cells were grown aerobically at 37 °C in LB supplemented with ampicillin (200 µg ml<sup>-1</sup>). Expression was induced at  $A_{600\text{ nm}} = 0.6$  with 1 mM IPTG for 3 h. Protein extraction was performed as described for Trx and purification was performed as described for MsrP-His<sub>6</sub>.

**In vitro repair of oxidized CaM, SurA and Pal by MsrP.** CaM was oxidized *in vitro* as described previously<sup>36</sup>. SurA-His<sub>6</sub> and Pal were oxidized *in vitro* by incubating the purified proteins (50 µM) for 2 h 30 min at 30 °C with 100 mM H<sub>2</sub>O<sub>2</sub> in a buffer containing 50 mM NaPi, pH 8.0, 50 mM NaCl. H<sub>2</sub>O<sub>2</sub> was then removed by gel filtration using a NAP-5 column (GE Healthcare) equilibrated with 50 mM NaPi, pH 8.0, 150 mM NaCl.

*In vitro* repair of oxidized CaM (CaMox), SurA (SurA ox) and Pal (Pal ox) was assessed by incubating the oxidized proteins (2 µM of CaMox and SurA ox, 5 µM of Pal ox) with purified MsrP-His<sub>6</sub> (2 µM for CaMox and SurA ox, 5 µM for Pal ox), 10 mM benzyl viologen and an excess of sodium dithionite at 37 °C for 1 h. As controls, the oxidized proteins were incubated separately with either MsrP-His<sub>6</sub> or the inorganic reducing system (benzyl viologen and sodium dithionite). The reactions were stopped by adding SB buffer and heating at 95 °C for the CaM and SurA samples or by adding 0.1% trifluoroacetic acid for the Pal samples. The CaM and SurA samples were then loaded on an SDS-PAGE gel and the proteins visualized with the PageBlue Protein Staining Solution (Fermentas). For the Pal samples (20 µg), proteins were separated by reverse-phase high-performance liquid chromatography on a C4 column (Vydac 214TP54, 4.6 mm × 250 mm) at a flow rate of 400 µl min<sup>-1</sup> with a linear gradient of acetonitrile in 0.1% trifluoroacetic acid (0–70% acetonitrile in 90 min). Absorbance was monitored at 214 nm and the peaks were collected. The fractions were dried in a Speedvac and the proteins resuspended in 25 µl of 100 mM NH<sub>4</sub>HCO<sub>3</sub> before overnight digestion at 30 °C with 0.5 µg of trypsin or EndoGlu-C. The peptides were then analysed as described below.

For CaM and SurA, the gel bands corresponding to the different oxidation states were in-gel digested with trypsin and the resulting peptides analysed by LC-MS/MS on a C18 reverse-phase column as described above. Relative abundances of every Met-containing peptide in its different oxidation state were obtained by integration

of peak area intensities, taking into account the extracted ion chromatogram of both doubly and triply charged ions.

**In vivo repair of oxidized SurA and Pal by MsrP.** The *in vivo* repair of SurA ox and Pal ox by the MsrPQ system or MsrP alone expressed from plasmids pAG195 and pAG192, respectively, was performed as follows. Overnight cultures of AG233 (containing the empty pAG177 vector), AG234 (containing the pAG195 plasmid) and AG289 (containing the pAG192 plasmid) were diluted to  $A_{600\text{ nm}} = 0.04$  into fresh LB medium (100 ml) and cells were grown aerobically at 37 °C in the presence of 0.1 mM IPTG and 200 µg ml<sup>-1</sup> ampicillin. At  $A_{600\text{ nm}} = 0.5$ , cells were subjected to NaOCl treatment (3.5 mM) and protein synthesis was blocked by the addition of chloramphenicol (300 µg ml<sup>-1</sup>). Samples were taken at different time points after NaOCl addition and precipitated with TCA. The pellets were then washed with ice-cold acetone, suspended in SB buffer, heated at 95 °C and loaded on a SDS-PAGE gel for immunoblot analysis using anti-Pal<sup>37</sup> and anti-SurA antibodies. The specificity of the anti-SurA antibody was verified (Supplementary Fig. 6). The protein amounts loaded were standardized by taking into account the  $A_{600\text{ nm}}$  values of the cultures.

**Oxidation, repair and purification of SurA for analysis of chaperone function.** SurA-Strep was oxidized *in vitro* by incubating the purified protein (200 µM) for 3 h at 30 °C with 100 mM H<sub>2</sub>O<sub>2</sub> in a buffer containing 50 mM NaPi, pH 8.0, 150 mM NaCl. H<sub>2</sub>O<sub>2</sub> was then removed by gel filtration using a NAP-5 column (GE Healthcare) equilibrated with 50 mM NaPi, pH 8.0, 150 mM NaCl. For the *in vitro* repair of oxidized SurA (SurA ox), the oxidized protein (30 µM) was incubated with purified MsrP-His<sub>6</sub> (30 µM), 10 mM benzyl viologen and 10 mM of sodium dithionite at 37 °C for 1 h. Following repair, SurA was purified by passing the sample through a gravity flow column containing 200 µl Strep-Tactin Sepharose beads (from a 50% suspension, IBA), previously equilibrated with buffer A (Tris-HCl 100 mM, pH 8.0, NaCl 150 mM, EDTA 1 mM). After washing with buffer A, repaired SurA was eluted using buffer A containing 2.5 mM desthiobiotin. The elution fractions were pooled and submitted to buffer exchange using a NAP-5 column (GE Healthcare) equilibrated with 50 mM NaPi, pH 8.0, 150 mM NaCl. To check for the correct oxidation, repair and purification of SurA, samples were loaded on an SDS-PAGE gel and the proteins visualized with the PageBlue Protein Staining Solution (Fermentas).

**Analysis of chaperone function.** The ability of SurA to act as a chaperone preventing the thermal aggregation of citrate synthase (Sigma, reference C3260) was assessed as follows. The aggregation of citrate synthase (0.15 µM) was monitored at 43 °C in 40 mM HEPES-KOH, pH 7.5, in the absence or in the presence of 0.6 µM SurA, SurA ox or MsrP-repaired SurA ox using light-scattering measurements. To avoid effects that might have been caused by the protein buffer, all samples were added to the assay in constant volume. SurA ox and MsrP-repaired SurA ox were obtained as described above. Light-scattering measurements were made using a Varian Cary Eclipse spectrofluorometer both with excitation and with emission wavelengths set to 500 nm at a spectral bandwidth of 2.5 nm. Data points were recorded every 0.1 s.

**Genetic analysis of Met-O assimilation.** The ability of various *E. coli* strains (BE100, JB08, CH193, BE104) to assimilate Met-O was assessed on M9 minimal medium supplemented with either Met or Met-O at 20 µg ml<sup>-1</sup>. Plates were incubated at 37 °C for 72 h. Overnight cultures of strains AG272, AG273, AG279 and AG274 were diluted to  $A_{600\text{ nm}} = 0.04$  into fresh M63 minimal medium (100 ml) supplemented with 0.5% glycerol, 150 µg ml<sup>-1</sup> of each amino acid, 1 mM MgSO<sub>4</sub>, 1 mM MoNa<sub>2</sub>O<sub>4</sub>, 17 µM Fe<sub>2</sub>(SO<sub>4</sub>)<sub>3</sub>, vitamins (thiamine 10 µg ml<sup>-1</sup>, biotin 1 µg ml<sup>-1</sup>, riboflavin 10 µg ml<sup>-1</sup>, and nicotinamide 10 µg ml<sup>-1</sup>) and 100 µg ml<sup>-1</sup> spectinomycin, and grown aerobically at 37 °C. When  $A_{600\text{ nm}}$  reached 0.5, cells (5 ml) were washed three times with M63 medium containing 150 µg ml<sup>-1</sup> Met-O instead of methionine, and serially diluted in the same medium. Five microlitres of each dilution were then spotted on M63 plates containing either Met or Met-O at 150 µg ml<sup>-1</sup>, and plates were subsequently incubated at 37 °C for 40 h.

**HOCl induction assays.** The *msrP::lacZ*-containing strains (CH183, CH186 and CH187) were grown at 37 °C with shaking in M9 minimal medium. When cells reached  $A_{600\text{ nm}} \approx 0.2$ , cultures were split into two plastic tubes, one of them containing HOCl (200 µM). These tubes were then incubated with an inclination of 90° with shaking at 37 °C. After 30 min of incubation, 1 ml was harvested and the bacteria were resuspended in 1 ml of β-galactosidase buffer. Levels of β-galactosidase were measured as described<sup>38</sup>.

**HOCl survival assays.** NR744, NR745, CH0127 and AG190 cells were grown aerobically at 37 °C with shaking in 50 ml of LB medium in 500 ml flasks. When cells reached  $A_{600\text{ nm}} \approx 0.45$ , 5 ml samples were transferred to conical polypropylene centrifuge tubes (50 ml; Sarstedt) and HOCl (2 mM) was added. Cells were then incubated at 37 °C with shaking (150 r.p.m.) at 90° inclination. Samples were taken at various time points after stress, diluted in PBS buffer, spotted on LB agar and incubated at 37 °C for 16 h. Cell survival was determined by counting



colony-forming units (c.f.u.) per millilitre. The absolute c.f.u. at time-point 0 (used as 100%) was  $\sim 10^8$  cells per millilitre in all experiments. For strains CH194, CH196 and CH197, the same protocol was used with chloramphenicol ( $25 \mu\text{g ml}^{-1}$ ) and arabinose (0.2%) added to the cultures.

**SDS survival assays.** Cells (MG1655 and BE107) were grown at  $37^\circ\text{C}$  with shaking in 10 ml of LB (in 100 ml flasks). When cells reached  $A_{600\text{nm}} \approx 0.8$ , 5 ml samples were transferred to conical polypropylene centrifuge tubes (50 ml, Sarstedt) and HOCl (2 mM) was added. After 5 min of incubation, samples were taken and diluted in PBS buffer to  $\sim 2 \times 10^3$  cells per millilitre. Aliquots ( $100 \mu\text{l}$ ) were then spread on LB agar plates containing SDS (1%). Colonies were counted the next day.

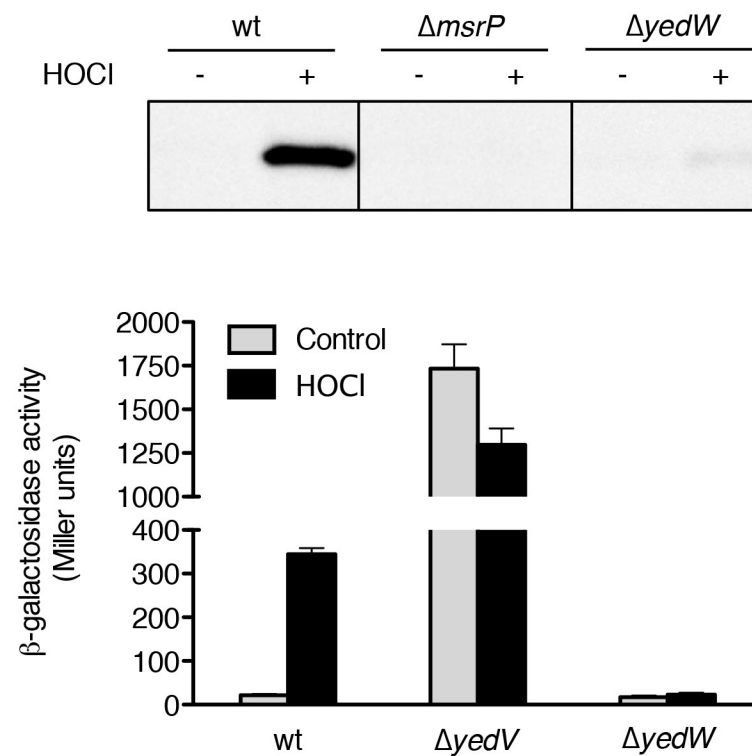
**Data set construction and phylogenetic analyses.** A non-redundant local protein database containing 1,342 complete prokaryotic proteomes available in NCBI (<http://www.ncbi.nlm.nih.gov/>) as of 30 July 2014 was built. This database was queried with the BlastP program (default parameters)<sup>39</sup>, using YedY (NP\_416480) and YedZ (NP\_416481) of *E. coli* strain K-12 substrate MG1655 as a seed. Distinction between homologous and non-homologous sequences was assessed by visual inspection of each BlastP output (no arbitrary cut-off on the *E* value or score). To ensure that we did not overlook divergent YedY or YedZ proteins, iterative BlastP queries were performed using homologues identified at each step as new seeds. The list of YedY and YedZ homologues is provided in Supplementary Data 1. The retrieved sequences were aligned using MAFFT version 7 (default parameters<sup>40</sup>; Supplementary Data 2 and 3). Each alignment was visually inspected and manually refined when necessary using the ED program from the MUST package<sup>41</sup>. Regions where the homology between amino-acid positions was doubtful were removed by using BMGE software (BLOSUM30 similarity matrix<sup>42</sup>).

For each homologue, the genomic context was investigated using MGcV (Microbial Genomic context Viewer<sup>43</sup>). The domain composition and protein location of each homologue was also analysed using pfam version 27.0 (ref. 44), SignalP version 4.1 (ref. 45) and TMHMM server version 2.0 (ref. 46), respectively.

For the YedY protein, preliminary phylogenetic analysis used FastTree version 2 and a gamma distribution with four categories<sup>47</sup>. On the basis of the resulting tree, the subfamily containing the sequence from *E. coli* was identified and selected for further phylogenetic investigations. The corresponding sequences were realigned using MAFFT version 7. The resulting alignment was trimmed with BMGE as previously described.

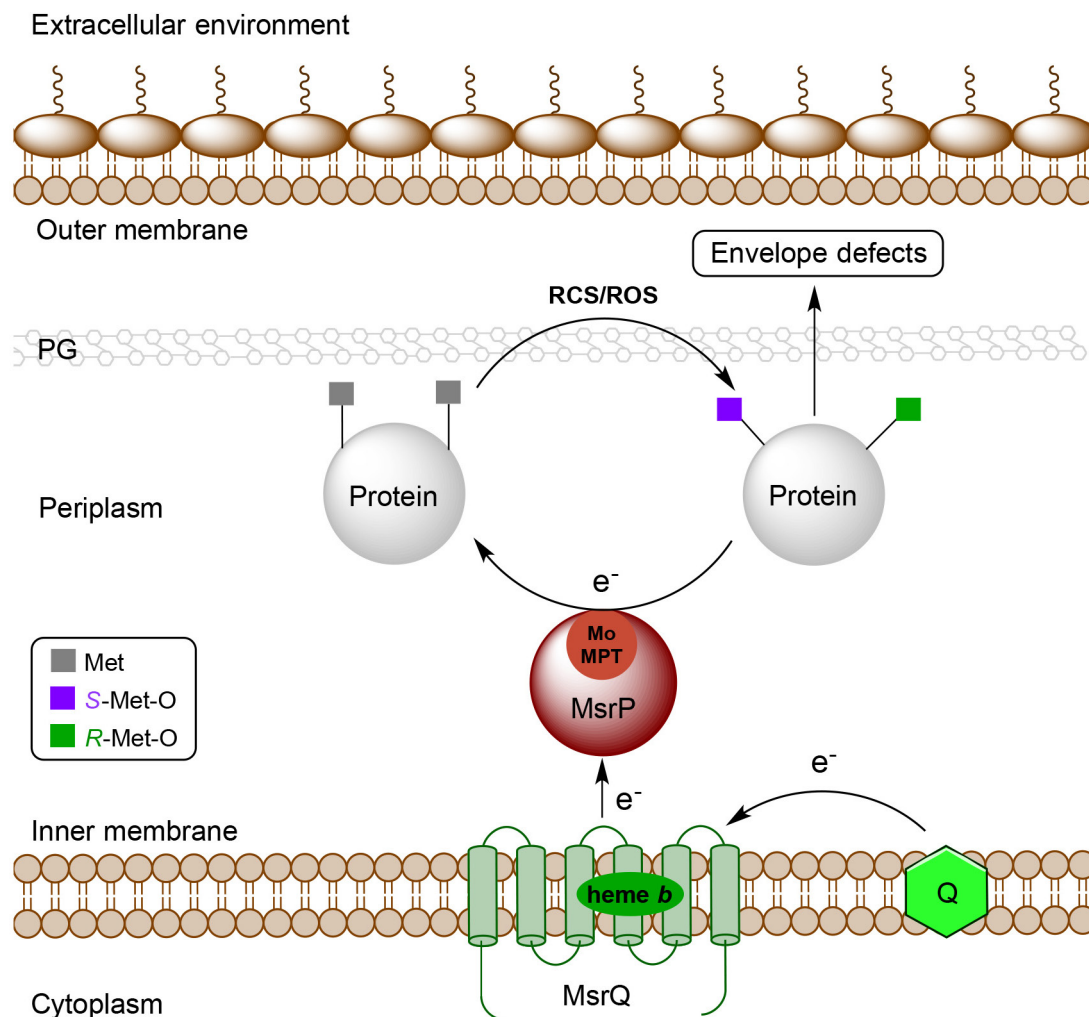
Maximum likelihood trees were computed using PHYML version 3.1 (ref. 48) with the Le and Gascuel model (amino-acid frequencies estimated from the data set) and a gamma distribution (four discrete categories of sites and an estimated alpha parameter) to take into account variations in evolutionary rate across sites. Branch robustness was estimated by the non-parametric bootstrap procedure implemented in PhyML (100 replicates of the original data set with the same parameters). Bayesian inferences were performed using MrBayes 3.2 (ref. 49) with a mixed model of amino-acid substitution including a gamma distribution (four discrete categories). MrBayes was run with four chains for one million generations and trees were sampled every 100 generations. To construct the consensus tree, the first 2,000 trees were discarded as 'burn in'.

22. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006.0008 (2006).
23. Bremer, E., Silhavy, T. J., Weisemann, J. M. & Weinstock, G. M. Lambda placMu: a transposable derivative of bacteriophage lambda for creating lacZ protein fusions in a single step. *J. Bacteriol.* **158**, 1084–1093 (1984).
24. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
25. Mandin, P. & Gottesman, S. A genetic approach for finding small RNAs regulators of genes of interest identifies RybC as regulating the DpiA/DpiB two-component system. *Mol. Microbiol.* **72**, 551–565 (2009).
26. Gottlieb, H. E., Kotlyar, V. & Nudelman, A. NMR chemical shifts of common laboratory solvents as trace impurities. *J. Org. Chem.* **62**, 7512–7515 (1997).
27. Holland, H. L., Andreana, P. R. & Brown, F. M. Biocatalytic and chemical routes to all the stereoisomers of methionine and ethionine sulfoxides. *Tetrahedron Asym.* **10**, 2833–2843 (1999).
28. Lavine, T. F. The formation, resolution, and optical properties of the diastereoisomeric sulfoxides derived from L-methionine. *J. Biol. Chem.* **169**, 477–491 (1947).
29. Koc, A., Gasch, A. P., Rutherford, J. C., Kim, H. Y. & Gladyshev, V. N. Methionine sulfoxide reductase regulation of yeast lifespan reveals reactive oxygen species-dependent and -independent components of aging. *Proc. Natl Acad. Sci. USA* **101**, 7999–8004 (2004).
30. Lherbet, C., Gravel, C. & Keillor, J. W. Synthesis of S-alkyl L-homocysteine analogues of glutathione and their kinetic studies with  $\gamma$ -glutamyl transpeptidase. *Bioorg. Med. Chem. Lett.* **14**, 3451–3455 (2004).
31. Vertommen, D. *et al.* The disulphide isomerase DsbC cooperates with the oxidase DsbA in a DsbD-independent manner. *Mol. Microbiol.* **67**, 336–349 (2008).
32. Arts, I. S. *et al.* Dissecting the machinery that introduces disulfide bonds in *Pseudomonas aeruginosa*. *MBio* **4**, e00912–13 (2013).
33. Vizcaino, J. A. *et al.* ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination. *Nature Biotechnol.* **30**, 223–226 (2004).
34. Roberts, D. M. *et al.* Chemical synthesis and expression of a calmodulin gene designed for site-specific mutagenesis. *Biochemistry* **24**, 5090–5098 (1985).
35. Grimaud, R. *et al.* Repair of oxidized proteins. Identification of a new methionine sulfoxide reductase. *J. Biol. Chem.* **276**, 48915–48920 (2001).
36. Tsvetkov, P. O. *et al.* Calorimetry and mass spectrometry study of oxidized calmodulin interaction with target and differential repair by methionine sulfoxide reductases. *Biochimie* **87**, 473–480 (2005).
37. Cascales, E., Bernadac, A., Gavioli, M., Lazzaroni, J. C. & Lloubes, R. Pal lipoprotein of *Escherichia coli* plays a major role in outer membrane integrity. *J. Bacteriol.* **184**, 754–759 (2002).
38. Miller, J. A. *Short Course in Bacterial Genetics* Unit 3, 72–74 (Cold Spring Harbor Laboratory Press, 1992).
39. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
40. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
41. Philippe, H. MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res.* **21**, 5264–5272 (1993).
42. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
43. Overmars, L., Kerkhoven, R., Siezen, R. J. & Francke, C. MGcV: the microbial genomic context viewer for comparative genome analysis. *BMC Genomics* **14**, 209 (2013).
44. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
45. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**, 785–786 (2011).
46. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182 (1998).
47. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
48. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
49. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).



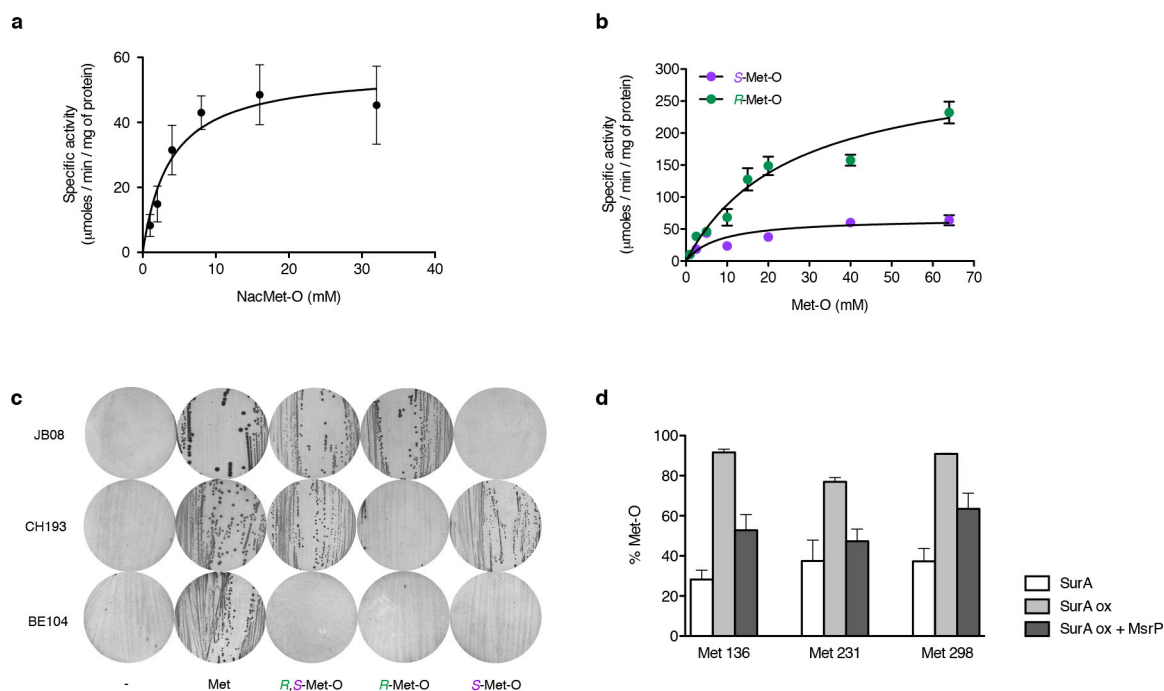
**Extended Data Figure 1 | Induction of MsrPQ by HOCl is dependent on the presence of a functional YedVW two-component system.** Top, immunoblot analysis shows that the induction of MsrP synthesis by HOCl (0.2 mM) is *yedW*-dependent. The image is representative of experiments

made in biological triplicate. Bottom, an *msrP::lacZ* fusion was used as a read-out for *msrP* expression. Deletion of *yedV* upregulates *msrP* expression, while deletion of *yedW* prevents its induction by HOCl. Error bars, mean ± s.e.m.; *n* = 4. The uncropped blot is shown in Supplementary Fig. 4.



**Extended Data Figure 2 | Respiratory chain-powered, non-stereospecific reduction of Met-O in periplasmic proteins by the MsrPQ system maintains envelope integrity.** Upon exposure to reactive species of chlorine (RCS) and/or reactive species of oxygen (ROS), methionine residues (Met) in periplasmic proteins such as SurA and Pal get oxidized and randomly form either the *R*- or the *S*- diastereoisomer of Met-O. This results in the loss of function of some proteins important

for maintaining the integrity of the envelope, such as SurA, giving rise to envelope defects. MsrP catalyses the reduction of both diastereoisomers of Met-O with the help of its molybdenum-molybdopterine (Mo-MPT) cofactor. Electrons for reduction are provided by the quinone (Q) pool of the respiratory chain through MsrQ, the inner membrane haem *b*-containing partner of MsrP. PG, peptidoglycan.

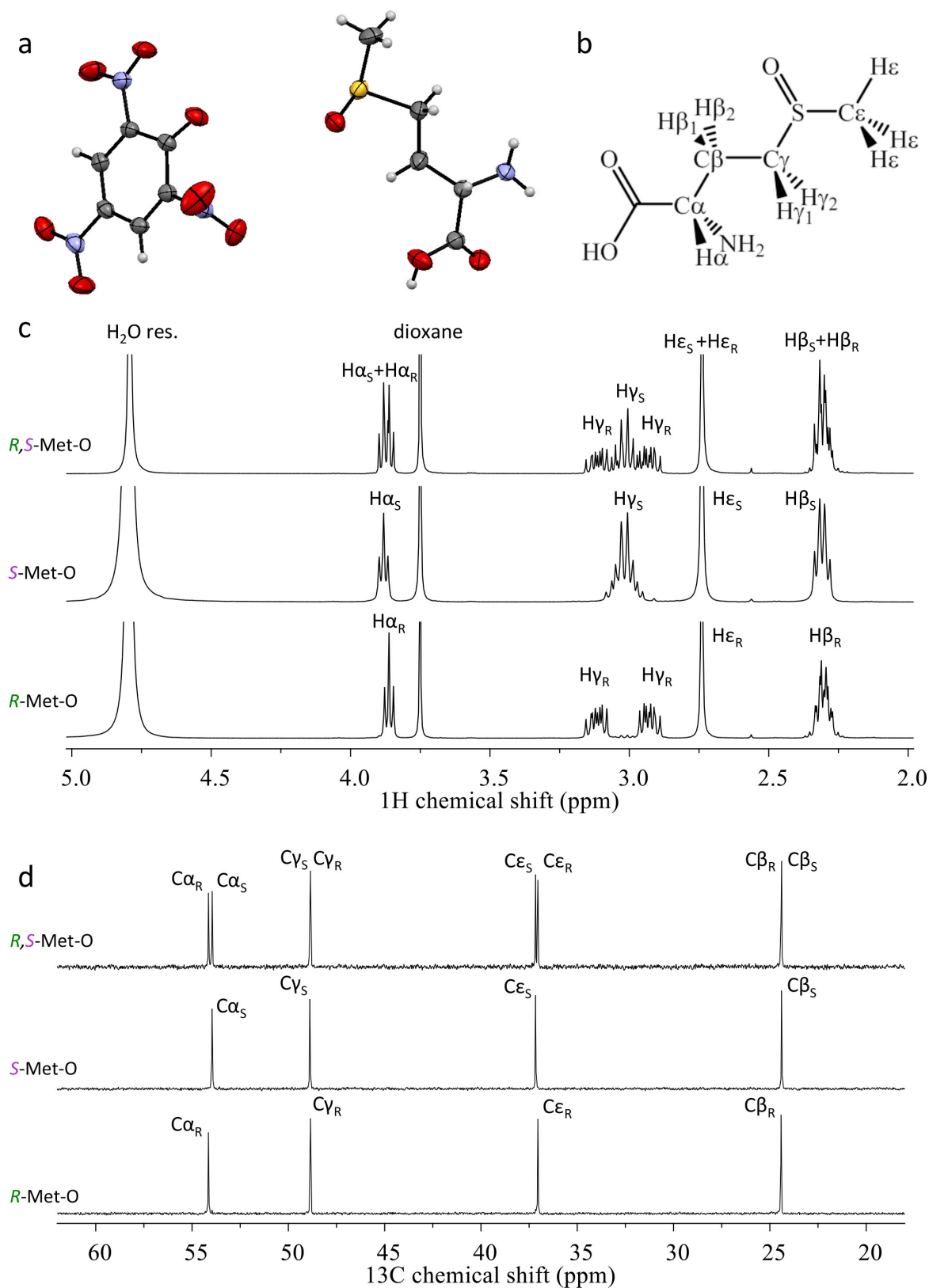


### Extended Data Figure 3 | MsrP non-stereospecifically reduces Met-O.

**a**, MsrP reduces *N*-acetyl-Met-O (NacMet-O), a substrate mimicking protein-bound Met-O, with  $K_m = 3.8 \pm 1.2$  mM, turnover number ( $k_{\text{cat}}$ ) =  $30.5 \pm 3.1$  s<sup>-1</sup> and  $V_{\text{max}} = 56.3 \pm 5.8$   $\mu\text{mol min}^{-1}$  per milligram protein (error bars, mean  $\pm$  s.d.;  $n = 3$ ). **b**, MsrP is a non-stereospecific Msr, being able to reduce both *S*-Met-O (with  $K_m = 8.0 \pm 2.7$  mM,  $k_{\text{cat}} = 36.0 \pm 3.6$  s<sup>-1</sup> and  $V_{\text{max}} = 67.2 \pm 6.4$   $\mu\text{mol min}^{-1}$  per milligram protein) and *R*-Met-O (with  $K_m = 25.7 \pm 4.7$  mM,  $k_{\text{cat}} = 168.3 \pm 15.0$  s<sup>-1</sup> and  $V_{\text{max}} = 313.4 \pm 27.6$   $\mu\text{mol min}^{-1}$  per milligram protein). Error bars, mean  $\pm$  s.d.;  $n = 3$ . **c**, Strain JB08 (Met<sup>-</sup> MsrA<sup>-</sup> MsrB<sup>-</sup> BisC<sup>-</sup>, producing MsrC) is able to grow only on *R*-Met-O, whereas strain CH193

(Met<sup>-</sup> MsrA<sup>-</sup> MsrB<sup>-</sup> MsrC<sup>-</sup>, producing BisC) is only able to grow on *S*-Met-O. Deletion of *msrP* in strain BE100 (Met<sup>-</sup> Msr<sup>-</sup> Sup<sup>Met-O+</sup>) prevents its growth on *R*- and *S*-Met-O (strain BE104 = Met<sup>-</sup> Msr<sup>-</sup> Sup<sup>Met-O+</sup>  $\Delta$  *msrP*, compare with growth of BE100 in Fig. 2e). Images are representative of experiments made in biological triplicate. **d**, The periplasmic chaperone SurA was treated with H<sub>2</sub>O<sub>2</sub>, giving rise to SurA ox, a sample of which was subsequently incubated with MsrP and the inorganic reducing system *in vitro*. The oxidation state of specific Met residues (Met 136, 231 and 298) in the various samples was determined by LC-MS/MS analysis. Error bars, mean  $\pm$  s.e.m.;  $n = 4$ .





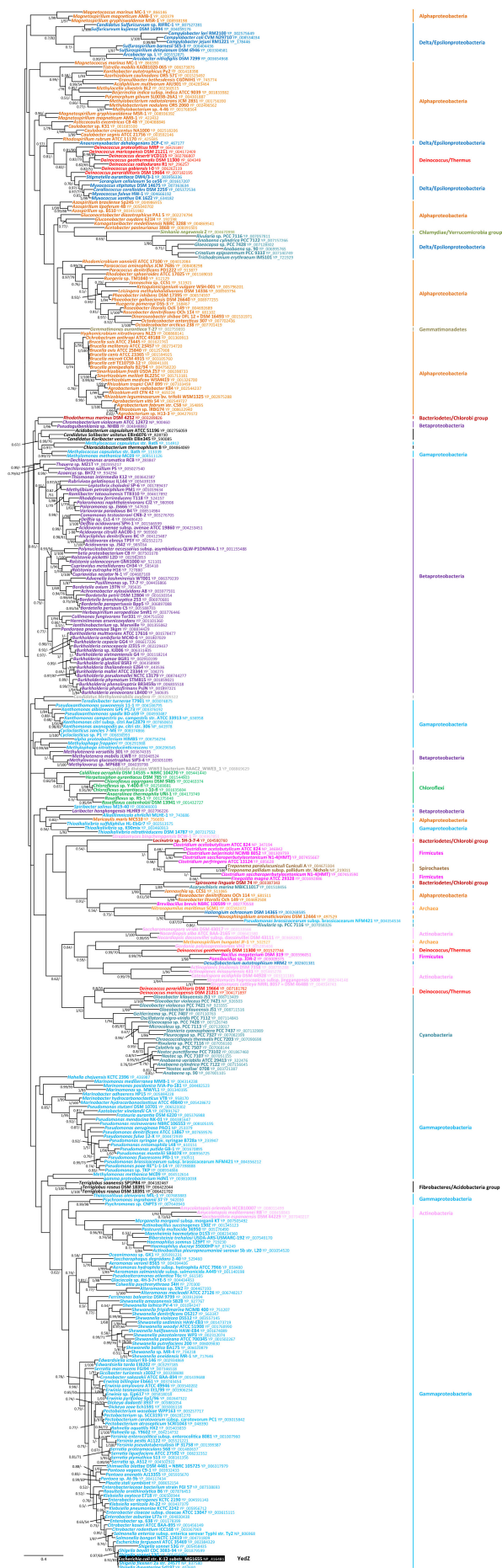
**Extended Data Figure 4 | Preparation of pure diastereoisomeric forms of Met-O.** **a**, The Oak Ridge Thermal Ellipsoid Plot (ORTEP ellipsoid) representation with 50% probability level of the crystal structure for the isolated salt of L-methionine-S-sulfoxide (right) picrate (left). The grey, blue, red, white and yellow spheres respectively represent carbon, nitrogen, oxygen, hydrogen and sulfur atoms. **b**, Chemdraw representation of L-methionine-R,S-sulfoxide with proton and carbon positioning (relative to NMR assignment). **c**, Zoom on the  $^1\text{H}$  NMR spectra of  $\sim 150$  mM solutions

of L-methionine sulfoxide in  $\text{D}_2\text{O}$  pD 6.5, either as a mixture of R- and S- diastereoisomers (top), isolated S- (middle) or isolated R- (bottom) (containing 30 mM dioxane as an internal reference). **d**, Zoom on the  $^{13}\text{C}$  NMR spectra of  $\sim 150$  mM solutions of L-methionine sulfoxide in  $\text{D}_2\text{O}$  pD 6.5, either as a mixture of R- and S- diastereoisomers (top), isolated S- (middle) or isolated R- (bottom) (containing 30 mM dioxane as an internal reference).



**Extended Data Figure 5 | Individual phylogenies of YedY.** Shown are unrooted Bayesian phylogenetic trees for YedY (b1971, 310 sequences, 260 positions). Numbers at nodes indicate posterior probabilities computed by MrBayes<sup>49</sup> and bootstrap values computed by PhyML<sup>48</sup>.

Only posterior probabilities and bootstrap values above 0.5 and 50%, respectively, are shown. Scale bars, average number of substitutions per site. In the phylogenetic tree, YedY from *E. coli* is highlighted in grey.



Alphaproteobacteria

Delta/Episymbiontobacteria

Alphaproteobacteria

Delta/Episymbiontobacteria

Deinococcus/Thermus

Delta/Episymbiontobacteria

Alphaproteobacteria

Chlamydiae/Tenericutes group

Delta/Episymbiontobacteria

Alphaproteobacteria

Gammaproteobacteria

Bacteroidetes/Chlorobi group

Betaproteobacteria

Gammaproteobacteria

Betaproteobacteria

Gammaproteobacteria

Betaproteobacteria

Chloroflexi

Betaproteobacteria

Alphaproteobacteria

Gammaproteobacteria

Bacteroidetes/Chlorobi group

Firmicutes

Spirichaeia

Bacteroidetes/Chlorobi group

Alphaproteobacteria

Archaea

Actinobacteria

Acidobacteria

Deinococcus/Thermus

Firmicutes

Actinobacteria

Deinococcus/Thermus

Cyanobacteria

Gammaproteobacteria

Fibrobacteres/Actinobacteria group

Gammaproteobacteria

Actinobacteria

Gammaproteobacteria



**Extended Data Figure 6 | Individual phylogenies of YedZ.** Shown are unrooted Bayesian phylogenetic trees for YedZ (b1972, 369 sequences, 135 positions). Numbers at nodes indicate posterior probabilities computed by MrBayes<sup>49</sup> and bootstrap values computed by PhyML<sup>48</sup>.

Only posterior probabilities and bootstrap values above 0.5 and 50%, respectively, are shown. Scale bars, average number of substitutions per site. In the phylogenetic tree, YedZ from *E. coli* is highlighted in grey.

**Extended Data Table 1 | The MsrPQ system uses electrons from the respiratory chain to reduce free Met-O**

Strain description	Met	Met-O
Met	+	+
Met Msr <sup>-</sup> (JB590)	+	-
Met Msr <sup>-</sup> Sup <sup>Met-O<sup>+</sup></sup> (BE100)	+	+
Met Msr <sup>-</sup> Sup <sup>Met-O<sup>+</sup></sup> $\Delta yedZ$ (BE105)	+	-
Met Msr <sup>-</sup> empty vector (AG272)	+	-
Met Msr <sup>-</sup> <i>pyedY</i> (AG273)	+	-
Met Msr <sup>-</sup> <i>pyedZ</i> (AG279)	+	-
Met Msr <sup>-</sup> <i>pyedYyedZ</i> (AG274)	+	+
Met Msr <sup>-</sup> Sup <sup>Met-O<sup>+</sup></sup> $\Delta menA \Delta ubiE$ (BE106)	+	-

This table shows the ability of the various strains to grow (+) or not (-) using Met-O as the sole Met source. Strains were grown for 40–72 h at 37 °C. The results are representative of experiments made in biological triplicate.

Extended Data Table 2 | List of proteins identified as potential MsrP substrates

Protein	Function	Number and percentage of methionines in the protein*
SurA	Primary periplasmic chaperone	14 (3.4%)
LolA	Outer-membrane lipoprotein carrier protein	2 (1.1%)
Pal	Peptidoglycan-associated lipoprotein	6 (3.9%)
MlaC	Probable phospholipid-binding protein	4 (2.1%)
PpiA	Peptidyl-prolyl cis-trans isomerase A	4 (2.4%)
DsbA	Thiol:disulfide interchange protein	6 (3.2%)
CysP	Thiosulfate-binding protein	4 (1.3%)
PotD	Spermidine/putrescine-binding periplasmic protein	9 (2.8%)
MppA	Periplasmic murein peptide-binding protein	7 (1.4%)
ProX	Glycine betaine-binding periplasmic protein	6 (1.9%)
MalE	Maltose-binding periplasmic protein	6 (1.6%)
MglB	D-galactose-binding periplasmic protein	6 (1.9%)
RbsB	D-ribose-binding periplasmic protein	4 (1.5%)
FecB	Fe <sup>3+</sup> dicitrate-binding periplasmic protein	7 (2.5%)
RcnB	Nickel/cobalt homeostasis protein	2 (2.3%)
ZnuA	High-affinity zinc uptake system protein	6 (2.1%)
Ecotin	General inhibitor of pancreatic serine proteases	4 (2.8%)
Ivy	Inhibitor of vertebrate lysozyme	5 (3.9%)
PspE	Thiosulfate sulfurtransferase	2 (2.4%)
YmgD	Uncharacterized protein	4 (4.4%)

\*Referring to the mature protein without its signal sequence

Semi-quantitative two-dimensional LC-MS/MS analysis was used to identify proteins that have one or more oxidized Met residues that MsrP could reduce. The first column indicates the name of the protein, the second describes its function and the third gives the number and percentage of methionine residues in the mature protein (excluding the signal sequence).

# Neutrophils support lung colonization of metastasis-initiating breast cancer cells

Stefanie K. Wculek<sup>1</sup> & Ilaria Malanchi<sup>1</sup>

**Despite progress in the development of drugs that efficiently target cancer cells, treatments for metastatic tumours are often ineffective. The now well-established dependency of cancer cells on their microenvironment<sup>1</sup> suggests that targeting the non-cancer-cell component of the tumour might form a basis for the development of novel therapeutic approaches. However, the as-yet poorly characterized contribution of host responses during tumour growth and metastatic progression represents a limitation to exploiting this approach. Here we identify neutrophils as the main component and driver of metastatic establishment within the (pre-)metastatic lung microenvironment in mouse breast cancer models. Neutrophils have a fundamental role in inflammatory responses and their contribution to tumorigenesis is still controversial<sup>2–4</sup>. Using various strategies to block neutrophil recruitment to the pre-metastatic site, we demonstrate that neutrophils specifically support metastatic initiation. Importantly, we find that neutrophil-derived leukotrienes aid the colonization of distant tissues by selectively expanding the sub-pool of cancer cells that retain high tumorigenic potential. Genetic or pharmacological inhibition of the leukotriene-generating enzyme arachidonate 5-lipoxygenase (Alox5) abrogates neutrophil pro-metastatic activity and consequently reduces metastasis. Our results reveal the efficacy of using targeted therapy against a specific tumour microenvironment component and indicate that neutrophil Alox5 inhibition may limit metastatic progression.**

In the presence of a growing tumour, subclinical changes in leukocyte composition at distant sites have been reported to favour metastatic growth<sup>5–7</sup>. Cancer cells within a tumour are heterogeneous and retain different tumorigenic potentials. Nonetheless, metastasis-initiating cells (MICs) depend on a favourable microenvironment to grow efficiently at the distant site<sup>8–10</sup>. We therefore reasoned that an altered presence of leukocytes within distant tissues of tumour-bearing hosts might influence specific subsets of disseminating cancer cells. We investigated this hypothesis using the lung metastatic MMTV-polyoma middle T antigen (PyMT) mammary tumour mouse model, which allows monitoring of the cell subpopulation functionally defined by a higher metastasis initiation ability (CD24<sup>+</sup>CD90<sup>+</sup> MICs)<sup>8</sup>.

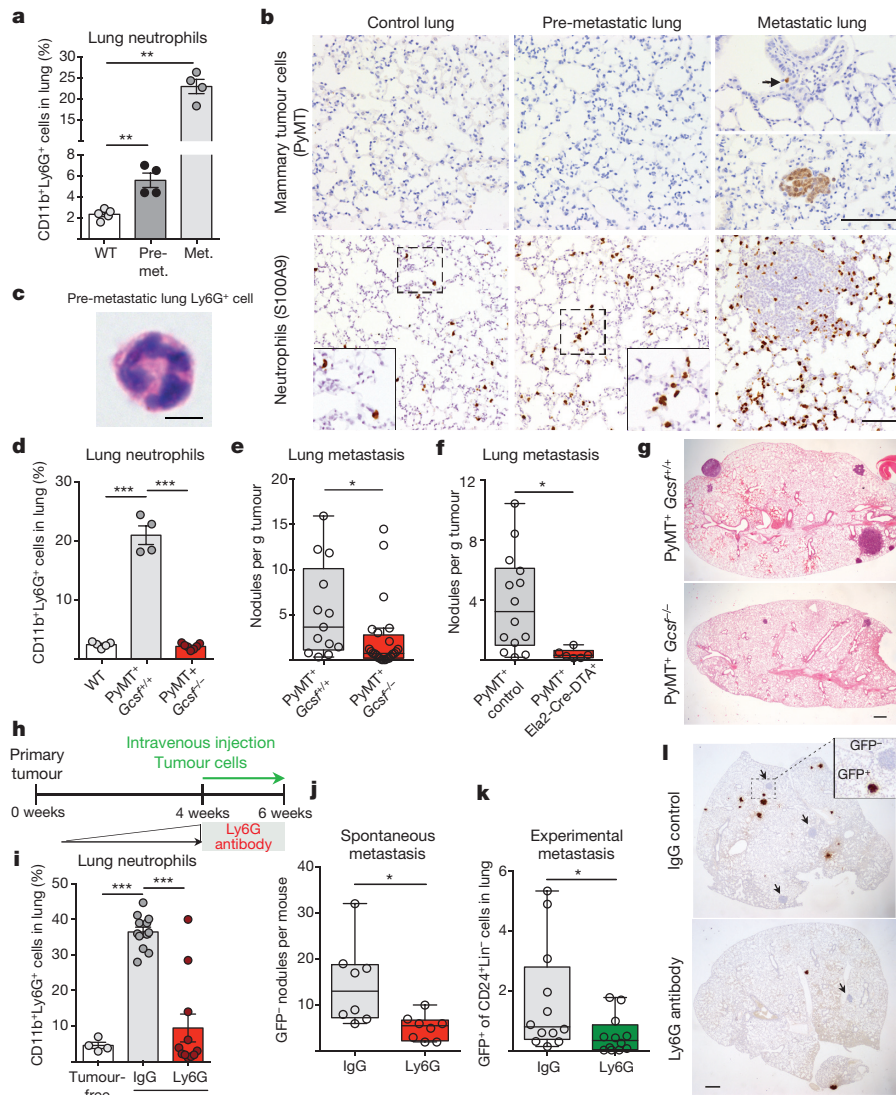
In accordance with previous reports<sup>11</sup>, we found CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophils to be systemically mobilized in MMTV-PyMT<sup>+</sup> tumour-bearing mice and, despite their low frequency within the primary tumour microenvironment, they were the main immune component that increased in metastatic lungs (Fig. 1a and Extended Data Fig. 1a–l). Importantly, CD11b<sup>+</sup>Ly6G<sup>+</sup> cells accumulated in the lung before cancer cells infiltrated the tissue (pre-metastatic lung) and their numbers increased during metastatic progression (metastatic lung) (Fig. 1a, b). We addressed the functional relevance of high CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophil numbers by analysing the metastatic progression of MMTV-PyMT<sup>+</sup> tumour-bearing mice in a neutropenic granulocyte colony-stimulating factor (*Gcsf*)-null background. Mice deficient in G-CSF expression developing mammary tumours failed to accumulate neutrophils in the lungs (Fig. 1d and Extended Data Fig. 2a). Notably, genetic neutropenia resulted in a robust reduction of spontaneous lung

metastasis, despite not affecting primary tumour growth (Fig. 1e, g and Extended Data Fig. 2b). No differences in lung macrophages compared with wild-type mice were detected (Extended Data Fig. 2c). Lack of G-CSF expression by cancer cells altered neither lung neutrophil accumulation nor metastasis (Extended Data Fig. 2d). In an alternative genetic strategy for neutrophil depletion, we crossed MMTV-PyMT<sup>+</sup> mice with neutrophil elastase (*Ela2*)-Cre and with ROSA-Flox-STOP-Flox diphtheria toxin (DTA) mice. Here, neutrophil-specific Cre expression led to DTA-mediated reduction of lung neutrophils in tumour-bearing mice, without altering lung macrophages and circulating myeloid cells or activating bone marrow natural killer (NK) and cytotoxic T cells (Extended Data Fig. 2e, f, h–j). Importantly, metastatic progression was impaired in MMTV-PyMT<sup>+</sup>-*Ela2*-Cre-DTA<sup>+</sup> mice without affecting primary tumour growth (Fig. 1f and Extended Data Fig. 2f, g).

Since lung neutrophil increase precedes cancer cell infiltration (Fig. 1b), we focused on the CD11b<sup>+</sup>Ly6G<sup>+</sup> cells accumulating in the early phase of lung colonization. We established mammary gland tumours by orthotopic transplantation to synchronize tumour growth, distant neutrophil accumulation and metastatic progression (Extended Data Fig. 3a). The comparison of tumour-induced CD11b<sup>+</sup>Ly6G<sup>+</sup> cells and CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophils from healthy lungs revealed minor variations, as messenger RNA expression of only two of seven tested neutrophil-secreted factors showed changes (Extended Data Fig. 3b). Tumour-mobilized lung neutrophils appeared morphologically mature (Fig. 1c) and the upregulation of CD31 suggests increased lung infiltration<sup>12</sup> (Extended Data Fig. 3b). Together, these data indicate that, at this time point, tumour-induced CD11b<sup>+</sup>Ly6G<sup>+</sup> cells in the lung are mature neutrophils similar to the ones found in healthy lungs. As neutrophils in the tumour context are reported to act as myeloid-derived suppressor cells<sup>13</sup>, we investigated the presence of an anticancer immune environment within the pre-metastatic lung of immune-competent mice. We used anti-Ly6G blocking antibody to deplete neutrophils during the pre-metastatic stage (Extended Data Fig. 4a). No significant differences were found in the frequencies and activation of various immune components as a consequence of neutrophil depletion, in particular in cytotoxic T and NK cells (Extended Data Figs 4b–o and 5a–i). To explore further the functional contribution of lung neutrophils to metastasis independently of potential immunosuppression, we performed time-controlled neutrophil depletion with anti-Ly6G antibody in immune-compromised mice (*Rag1*-null) harbouring primary tumours. Remarkably, pre-metastatic neutrophil depletion during metastatic colonization caused a decrease of spontaneous metastasis (Fig. 1h–j, l). Concomitantly, lungs of the same mice were synchronously seeded with cancer cells isolated from MMTV-PyMT<sup>+</sup> actin-green fluorescent protein (GFP) tumours by intravenous injection to initiate lung colonization (Fig. 1h). Notably, GFP<sup>+</sup> cancer cells colonizing neutrophil-depleted lungs were significantly reduced, revealing the relevance of lung neutrophils specifically during metastatic initiation (Fig. 1k, l). No alterations were found in the extravasation efficiency of labelled cancer cells (data not shown). Although we cannot exclude

<sup>1</sup>The Francis Crick Institute, Lincolns Inn Fields Laboratories, 44 Lincolns Inn Fields, London WC2A 3LY, UK.





**Figure 1 | Neutrophils infiltrate pre-metastatic lungs and favour metastasis.** **a, b,** Analysis of wild-type (WT) or MMTV-PyMT<sup>+</sup> mice. **a,** Lung neutrophils frequencies determined by flow cytometry ( $n = 5$  (wild type),  $n = 4$  (pre-metastatic lung),  $n = 4$  (metastatic lung)). **Met.,** metastatic. **b,** Lung neutrophils or cancer cells determined by histology staining for S100A9 or PyMT (brown). Scale bars, 100  $\mu\text{m}$ . Magnifications in inserts. **c,** Haematoxylin & eosin (H&E)-stained neutrophil. Scale bar, 5  $\mu\text{m}$ . **d,** Lung neutrophil quantification by flow cytometry ( $n = 5$  (wild type),  $n = 4$  (PyMT<sup>+</sup> *Gcsf*<sup>+/+</sup>),  $n = 7$  (PyMT<sup>+</sup> *Gcsf*<sup>-/-</sup>)). **e, f,** Spontaneous metastasis of MMTV-PyMT<sup>+</sup> *Gcsf*<sup>+/+</sup> ( $n = 13$ ) or MMTV-PyMT<sup>+</sup> *Gcsf*<sup>-/-</sup> ( $n = 24$ )

(**e**) and MMTV-PyMT<sup>+</sup> control ( $n = 14$ ) or MMTV-PyMT<sup>+</sup>Ela2-Cre-RTA<sup>+</sup> ( $n = 6$ ) mice (**f**). **g,** Representative H&E-stained sections of lung. Scale bar, 500  $\mu\text{m}$ . **h,** Experimental setup for neutrophil depletion. **i,** Flow cytometric lung neutrophil quantification ( $n = 4$  (tumour-free),  $n = 12$  (IgG tumour),  $n = 11$  (Ly6G tumour)). **j, k,** Spontaneous ( $n = 8$  per group) (**j**) and experimental metastasis ( $n = 12$  per group) (**k**). Lin, CD45, CD31, TER119. **l,** Histological GFP-stained lung sections including close-up on spontaneous (arrow) and experimental metastases (brown). Scale bar, 500  $\mu\text{m}$ . Statistical analysis by two-sided *t*-test. Data are represented as mean  $\pm$  standard error of the mean (s.e.m.). \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

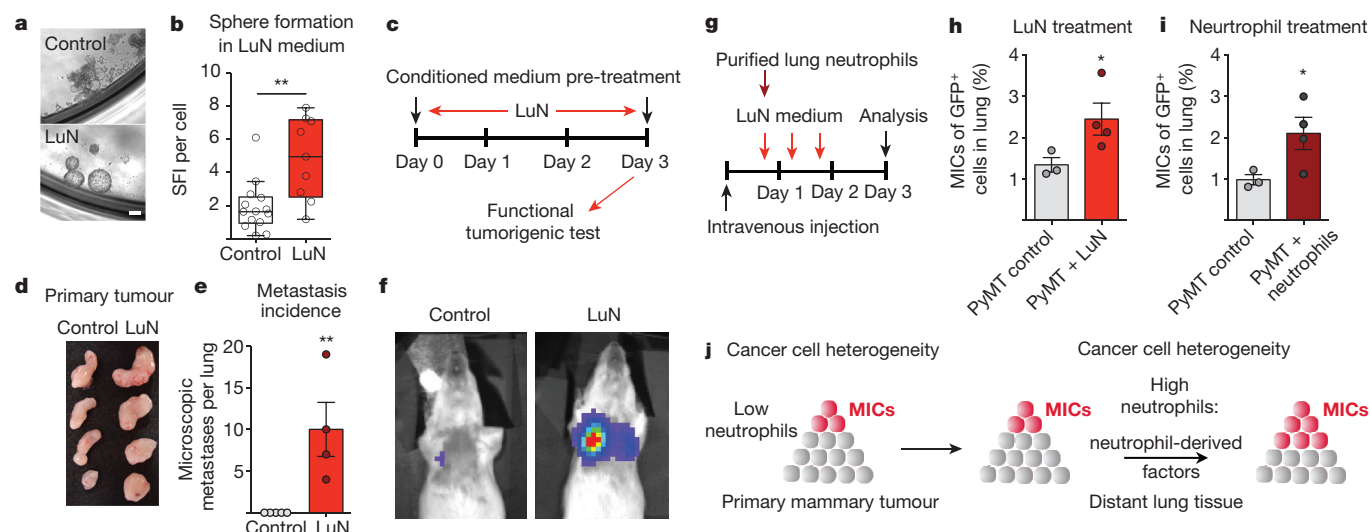
a contribution of other cells to a favourable pre-metastatic environment<sup>5–7</sup>, such as monocytes<sup>14</sup>, these results reveal that the breast-tumour-induced systemic accumulation of neutrophils coincidentally acts as a pre-metastatic niche in tissue targeted for metastatic dissemination.

Next, we investigated a potential direct effect of neutrophil-secreted factors on tumour cells. Pre-metastatic lung neutrophils (Extended Data Fig. 6a, b) were used to condition cell culture medium for 14 h (LuN medium). Primary MMTV-PyMT tumour cells cultured in LuN medium in non-adherent culture showed enhanced sphere growth (Fig. 2a, b). Furthermore, short-term exposure to LuN medium in adherent culture boosted the tumorigenic potential of cancer cells *in vivo* and *in vitro* (Fig. 2c, d and Extended Data Fig. 6c, d). Importantly, short-term culture in LuN medium also increased the metastatic initiation potential of total cancer cells (Fig. 2e, f).

Cancer cells are also heterogeneous when disseminated into the circulation<sup>15</sup> and might respond differently to environmental

stimulations<sup>16</sup>. We therefore probed whether neutrophil-secreted factors influence the relative amount of highly metastatic cells. We monitored the previously described MIC population (CD24<sup>+</sup>CD90<sup>+</sup>)<sup>8</sup> after exposing tumour cells seeded into the lung to either LuN medium or freshly isolated pre-metastatic lung neutrophils (Fig. 2g). Notably, both settings induced a doubling of MIC frequencies among the total cancer cell population (Fig. 2h, i and Extended Data Fig. 6e–h) and partially increased metastatic growth (Extended Data Fig. 6i–k). Collectively, we observe that neutrophil-derived factors alter the heterogeneity of cancer cells favouring MICs and lead to increased metastatic competence of total cancer cells (Fig. 2j).

We aimed to identify neutrophil-secreted factors mediating this activity. LuN medium contains many factors (data not shown) including CCL2, MMP9, interleukin (IL)-6 and IL-1 that might alter inflammatory responses and increase pro-tumorigenic behaviour<sup>17–19</sup>. Various cells in the tumour microenvironment can secrete these mediators,



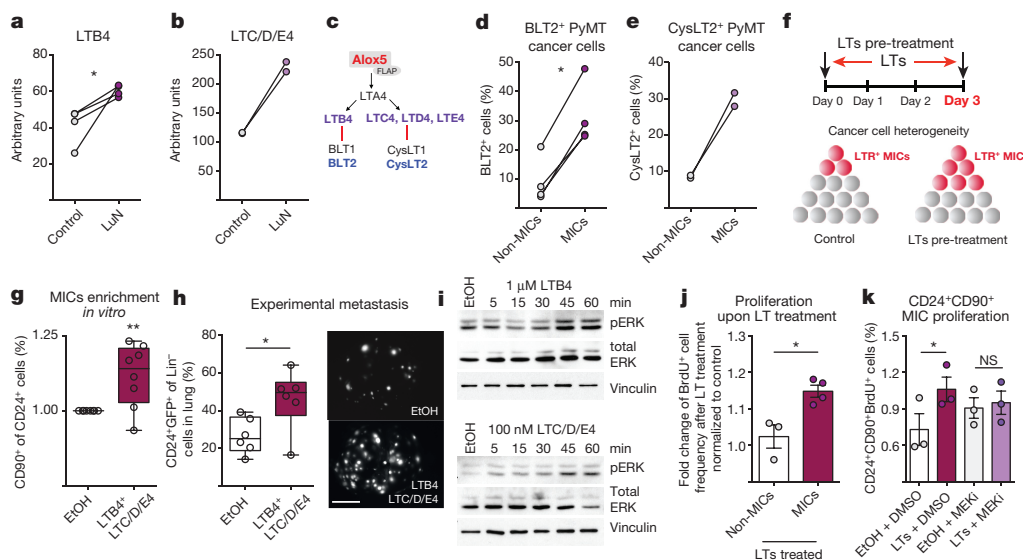
**Figure 2 | Neutrophil-derived signals promote tumorigenicity and increase the metastatic cell sub-pool.** **a, b**, Images and quantification (technical replicate  $n = 14$  (control),  $n = 9$  (LuN)) of biological triplicates of primary MMTV-PyMT spheres in indicated medium. SFI, sphere formation index. Scale bar,  $10\mu\text{m}$ . **c–f**, Medium pre-treated luciferase<sup>+</sup> MMTV-PyMT cells (**c**) grafted onto the mammary gland (**d**) or intravenously injected (**e, f**) into *Rag1*-null mice. Lung metastases quantified by histological sectioning ( $n = 5$  (control),  $n = 4$  (LuN)).

**f**, Representative bioluminescence signal. **g**, Experimental setup. **h, i**, Flow cytometric quantification of MICs in lungs of LuN-treated ( $n = 3$  (PyMT control),  $n = 4$  (PyMT + LuN)) (**h**) or neutrophil-treated mice ( $n = 3$  (PyMT control),  $n = 4$  (PyMT + neutrophils)) (**i**). **j**, Representation of cell heterogeneity change. Statistical analysis by two-sided *t*-test (**b**), Mann–Whitney test (**e**) and one representative experiment of two analysed by analysis of variance (ANOVA) (**h, i**). Data are represented as mean  $\pm$  s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$ .

so we concentrated on specific innate leukocyte-derived factors. We detected high levels of the lipids leukotriene B4 (LTB4) and cysteinyl leukotrienes C4, D4 and E4 (LTC/D/E4), products of the Alox5 enzyme<sup>20</sup> (Fig. 3a–c). Importantly, direct leukotriene (LT) stimulation boosted sphere formation and a short 3-day LT exposure of total cancer cells enhanced their tumour initiation potential (Extended Data Fig. 7a–c). Notably, cells expressing LT receptors (LTRs; LTB4 receptor 2 (BLT2) and LTC/E/D4 receptor 2 (CysLT2))<sup>21,22</sup> appeared to be enriched among MICs within total MMTV-PyMT cancer cells as well as among other known tumorigenic subpopulations of breast cancer

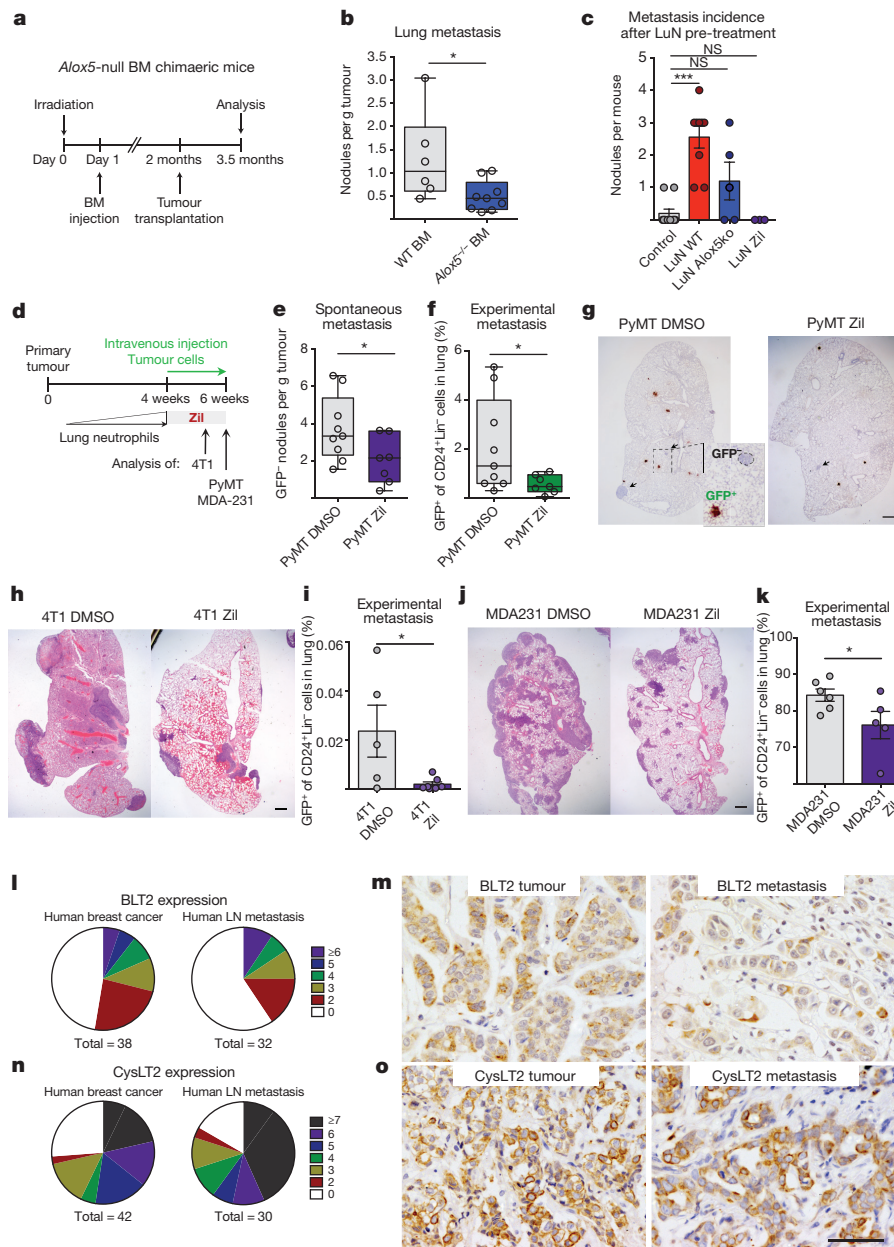
cell lines<sup>23–25</sup> (Fig. 3d, e and Extended Data Fig. 7d–i). Indeed, LTRs themselves identified MMTV-PyMT cancer cells with high sphere and tumour formation abilities (Extended Data Fig. 7j–l).

In accordance with LTR expression on MICs, we found that 3-day LT stimulation of MMTV-PyMT tumour cells *in vitro* increased MIC frequency and metastatic initiation capacity *in vivo* (Fig. 3f–h), similar to neutrophil-derived mediators (Fig. 2e–j). LT stimulation also enriched the CD49<sup>high</sup> sub-pool among 4T1 cells (Extended Data Fig. 8b). Other cells such as macrophages and eosinophils respond to LTs, but no broader inflammatory reaction was detected at this stage



**Figure 3 | LTs enrich for MICs and tumorigenicity.** **a, b**, Enzyme immunoassay detecting LTB4 ( $n = 4$  per group) (**a**) or LTC/D/E4 ( $n = 2$  per group) (**b**). **c**, Overview of LTs and LTRs. **d, e**, Flow cytometric quantification of BLT2<sup>+</sup> ( $n = 4$  tumours) (**d**) and CysLT2<sup>+</sup> cells ( $n = 2$  tumours) (**e**) among indicated sub-pools. **f–h**, Representation of LT treatment (**f**): frequency of MICs ( $n = 8$  per group) (**g**); and experimental lung metastasis ( $n = 6$  per group) with representative images of GFP<sup>+</sup> colonies (**h**). Scale bar, 3 mm. Lin, CD45, CD31, TER119. **i**, Western blot of ERK1/2 phosphorylation and

total ERK1/2 levels of LTB4- or LTC/D/E4-treated cells for indicated minutes. Loading control: anti-vinculin antibody. **j–k**, 5-Bromodeoxyuridine (BrdU) incorporation comparing LT-treated MICs with non-MICs ( $n = 3$  (non-MICs),  $n = 4$  (MICs)) (**j**) or MICs treated with LTs and/or PD0325901 MEK inhibitor (MEKi;  $n = 3$  per group) (**k**) DMSO, dimethylsulfoxide treated; EtOH, ethanol treated. Statistical analysis by two-sided *t*-test (**a, d, h, j, k**) and one-sample *t*-test (**g**). Data are represented as mean  $\pm$  s.e.m. NS, not significant. \* $P < 0.05$ , \*\* $P < 0.01$ . Blot source data are in Supplementary Fig. 1.



**Figure 4 | Alox5 inhibition decreases lung metastasis initiation.**

**a, b**, *Alox5*-null bone marrow (BM) chimaera experimental setup (**a**) and spontaneous metastasis ( $n = 6$  (wild-type bone marrow),  $n = 9$  (*Alox5*<sup>-/-</sup> bone marrow)) (**b**). WT, wild type. **c**, Surface metastases of medium pre-treated cancer cells ( $n = 10$  (control),  $n = 9$  (LuN wild type),  $n = 5$  (LuN *Alox5ko*),  $n = 3$  (LuN-Zil)). **d**, Experimental setup for Zil treatment. **e–k**, Spontaneous (**e**) and experimental (**f, i, k**) metastasis of MMTV-PyMT cells ( $n = 9$  (PyMT DMSO),  $n = 7$  (PyMT Zil)) (**e–g**), 4T1 cells ( $n = 5$  (4T1 DMSO),  $n = 7$  (4T1 Zil)) (**h, i**) or MDA-MB-231 cells ( $n = 6$

(MDA231 DMSO),  $n = 5$  (MDA231 Zil)) (**j, k**). Lin, CD45,CD31,TER119. Representative histological lung sections GFP stained with close-up on spontaneous (arrows) and experimental metastases (brown) (**g**) or H&E stained (**h, k**). Scale bars, 500  $\mu$ m. **l–o**, BLT2 (**l, m**) or CysLT2 (**n, o**) staining ( $n \geq 30$  per group). Quantification of staining intensity and frequency (**l, n**) and representative images (**m, o**). Scale bar, 50  $\mu$ m. Statistical analysis by two-sided *t*-test (**b, e, f, i**), Mann-Whitney test (**c**) and one-sided *t*-test (**k**). Data are represented as mean  $\pm$  s.e.m. NS, not significant, \* $P < 0.05$ , \*\*\* $P < 0.001$ .

(Extended Data Figs 4 and 5). In summary, LTs appear to shift heterogeneous cancer cell populations in favour of highly metastatic cells and enhance metastatic competence.

In line with previous reports on LTB<sub>4</sub> signalling<sup>21,26</sup>, cancer cells responded to both LTB<sub>4</sub> and LTC/D/E4 with increases in extracellular-signal-regulated kinases (ERK)1 and 2 phosphorylation (Fig. 3i and Extended Data Fig. 8c, d). LTR<sup>+</sup> cells were required to detect a LT-dependent phosphorylated (p)ERK1/2 increase (Extended Data Fig. 8e–g) and inhibitors for BLT2 and CysLT2 interfered with ERK1/2 activation (Extended Data Fig. 8h–k). Finally, 3-day LTC/D/E4 treatment increased the frequency of LTR<sup>+</sup> cancer cells, suggesting a functional boost in proliferation (Extended Data Fig. 8l). Indeed, LT

treatment specifically increased the proliferation of MICs in a MAPK/ERK kinases (MEK)1- and 2-mediated, pERK1/2-dependent manner (Fig. 3j, k and Extended Data Fig. 8m). These results indicate that LTs provide a selective proliferative advantage to cancer cells with intrinsically higher tumorigenicity (Extended Data Fig. 8a).

To confirm the functional relevance of LTs *in vivo*, we took advantage of an *Alox5*-null mouse model (Fig. 3c). We generated bone marrow chimaeric mice in which *Alox5* is genetically depleted in the radiosensitive immune cell compartment. Bone marrow *Alox5*-null mice grafted with MMTV-PyMT cells showed unaltered primary tumour growth and neutrophil lung accumulation (Fig. 4a and Extended Data Fig. 9a–d), yet the efficiency of spontaneous



metastasis was reduced (Fig. 4b). Next, we generated LT-deficient LuN (LuN-Alox5ko) medium from Alox5-null pre-metastatic lung neutrophils. Importantly, LuN-Alox5ko medium failed to boost the metastatic potential of luciferase-expressing MMTV-PyMT cells after 3-day pre-treatment (Figs 2c, 4c and Extended Data Fig. 9e, f). Taken together, these data confirm Alox5 products to be crucial for neutrophil pro-metastatic activity.

LTs are important mediators during inflammatory asthma and are targeted by the specific Alox5 inhibitor zileuton (Zil)<sup>27</sup>. We explored Zil-mediated inhibition of LT synthesis to treat metastatic breast cancer in mice. Zil blocked LT production *in vivo*, detected by decreased LTB4 levels in LuN medium (LuN-Zil) (Extended Data Fig. 10a, b) and, consequently, LuN-Zil medium failed to enhance metastasis (Fig. 4c). Importantly, in a therapeutic setting (Fig. 4d), treatment of MMTV-PyMT tumour-harboring mice with Zil reduced spontaneous metastasis (Fig. 4e, g), without altering primary tumours or lung neutrophil levels (Extended Data Fig. 10c, d). Additionally, the colonization capacity of GFP<sup>+</sup> MMTV-PyMT cancer cells seeded into lungs of Zil-treated mice was reduced (Fig. 4f, g). We confirmed that metastatic cancer cells showed reduced proliferation very early after infiltrating Zil-treated lungs (Extended Data Fig. 10e). Taken together, these data represent a potential therapeutic approach to target this novel LT/Alox5-dependent neutrophil pro-metastatic activity.

Importantly, similar results on the efficacy of Zil treatment in limiting metastatic progression were confirmed in two metastatic breast cancer cell lines, mouse 4T1 cells and human MDA-MB-231 cells (Fig. 4h–k and Extended Data Fig. 10f–i). As Zil treatment had no effect on long-term primary tumour growth *in vivo* or on cancer cell behaviour *in vitro* (Extended Data Fig. 10j–m), we exclude involvement of Alox5 products in a cancer-cell autocrine loop.

Clinical data correlating high neutrophil levels with poorer prognosis<sup>28,29</sup>, together with detected LTR expression in human metastatic ductal and lobular breast carcinoma and their lymph-node metastases (Fig. 4l–o), suggests that a similar neutrophil pro-metastatic mechanism might boost human breast cancer progression to the lung.

We have identified a novel LT/Alox5-dependent pro-metastatic activity of neutrophils supporting highly metastatic cells that can be targeted by Zil, offering hope for new cancer therapeutics.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 21 July; accepted 9 October 2015.**

**Published online 9 December 2015.**

1. Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nature Med.* **19**, 1423–1437 (2013).
2. Bald, T. *et al.* Ultraviolet-radiation-induced inflammation promotes angiogenesis and metastasis in melanoma. *Nature* **507**, 109–113 (2014).
3. Galdiero, M. R. *et al.* Tumor associated macrophages and neutrophils in cancer. *Immunobiology* **218**, 1402–1410 (2013).
4. Finisguerra, V. *et al.* MET is required for the recruitment of anti-tumoural neutrophils. *Nature* **522**, 349–353 (2015).
5. Hiratsuka, S., Watanabe, A., Aburatani, H. & Maru, Y. Tumour-mediated upregulation of chemoattractants and recruitment of myeloid cells predetermines lung metastasis. *Nature Cell Biol.* **8**, 1369–1375 (2006).
6. Erler, J. T. *et al.* Hypoxia-induced lysyl oxidase is a critical mediator of bone marrow cell recruitment to form the premetastatic niche. *Cancer Cell* **15**, 35–44 (2009).
7. Kaplan, R. N. *et al.* VEGFR1-positive haematopoietic bone marrow progenitors initiate the pre-metastatic niche. *Nature* **438**, 820–827 (2005).
8. Malanchi, I. *et al.* Interactions between cancer stem cells and their niche govern metastatic colonization. *Nature* **481**, 85–89 (2012).
9. Calon, A. *et al.* Dependency of colorectal cancer on a TGF- $\beta$ -driven program in stromal cells for metastasis initiation. *Cancer Cell* **22**, 571–584 (2012).

10. Oskarsson, T., Batlle, E. & Massagué, J. Metastatic stem cells: sources, niches, and vital pathways. *Cell Stem Cell* **14**, 306–321 (2014).
11. Coffelt, S. B. *et al.* IL-17-producing  $\gamma\delta$  T cells and neutrophils conspire to promote breast cancer metastasis. *Nature* **522**, 345–348 (2015).
12. Luu, N. T., Rainger, G. E., Buckley, C. D. & Nash, G. B. CD31 regulates direction and rate of neutrophil migration over and under endothelial cells. *J. Vasc. Res.* **40**, 467–479 (2003).
13. Condamine, T., Ramachandran, I., Youn, J. I. & Gabrilovich, D. I. Regulation of tumor metastasis by myeloid-derived suppressor cells. *Annu. Rev. Med.* **66**, 97–110 (2015).
14. Qian, B. Z. *et al.* CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis. *Nature* **475**, 222–225 (2011).
15. Yu, M. *et al.* Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* **339**, 580–584 (2013).
16. Korkaya, H., Liu, S. & Wicha, M. S. Breast cancer stem cells, cytokine networks, and the tumor microenvironment. *J. Clin. Invest.* **121**, 3804–3809 (2011).
17. Tsuyada, A. *et al.* CCL2 mediates cross-talk between cancer cells and stromal fibroblasts that regulates breast cancer stem cells. *Cancer Res.* **72**, 2768–2779 (2012).
18. Yan, H. H. *et al.* Gr-1<sup>+</sup>CD11b<sup>+</sup> myeloid cells tip the balance of immune protection to tumor promotion in the premetastatic lung. *Cancer Res.* **70**, 6139–6149 (2010).
19. Snoussi, K., Strosberg, A. D., Bouaouina, N., Ben Ahmed, S. & Chouchane, L. Genetic variation in pro-inflammatory cytokines (interleukin-1 $\beta$ , interleukin-1 $\alpha$  and interleukin-6) associated with the aggressive forms, survival, and relapse prediction of breast carcinoma. *Eur. Cytokine Netw.* **16**, 253–260 (2005).
20. Wang, D. & Dubois, R. N. Eicosanoids and cancer. *Nature Rev. Cancer* **10**, 181–193 (2010).
21. Cho, N. K., Joo, Y. C., Wei, J. D., Park, J. I. & Kim, J. H. BLT2 is a pro-tumorigenic mediator during cancer progression and a therapeutic target for anti-cancer drug development. *Am. J. Cancer Res.* **3**, 347–355 (2013).
22. Kanaoka, Y. & Boyce, J. A. Cysteinyl leukotrienes and their receptors: cellular distribution and function in immune and inflammatory responses. *J. Immunol.* **173**, 1503–1510 (2004).
23. Hiraga, T., Ito, S. & Nakamura, H. Side population in MDA-MB-231 human breast cancer cells exhibits cancer stem cell-like properties without higher bone-metastatic potential. *Oncol. Rep.* **25**, 289–296 (2011).
24. Sheridan, C. *et al.* CD44<sup>+</sup>/CD24<sup>−</sup> breast cancer cells exhibit enhanced invasive properties: an early step necessary for metastasis. *Breast Cancer Res.* **8**, R59 (2006).
25. Lo, P. K. *et al.* CD49f and CD61 identify Her2/neu-induced mammary tumor-initiating cells that are potentially derived from luminal progenitors and maintained by the integrin-TGF $\beta$  signaling. *Oncogene* **31**, 2614–2626 (2012).
26. Park, M. K. *et al.* Novel involvement of leukotriene B<sub>4</sub> receptor 2 through ERK activation by PP2A down-regulation in leukotriene B<sub>4</sub>-induced keratin phosphorylation and reorganization of pancreatic cancer cells. *Biochim. Biophys. Acta* **1823**, 2120–2129 (2012).
27. Wenzel, S. E. & Kamada, A. K. Zileuton: the first 5-lipoxygenase inhibitor for the treatment of asthma. *Ann. Pharmacother.* **30**, 858–864 (1996).
28. Donskov, F. Immunomonitoring and prognostic relevance of neutrophils in clinical trials. *Semin. Cancer Biol.* **23**, 200–207 (2013).
29. Han, Y. *et al.* Prognostic value of chemotherapy-induced neutropenia in early-stage breast cancer. *Breast Cancer Res. Treat.* **131**, 483–490 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank C. Reis e Sousa, E. Sahai, P. Scaffidi and J. Huelsken for scientific discussions, critical reading of the manuscript and sharing cell lines and mouse strains. We also thank members of the tumour-stroma interactions in cancer development (TSI) laboratory of The Crick Institute for scientific discussions, critical reading of the manuscript and practical support. We thank L. Jones for help in analysing the human breast cancer samples. We are grateful to R. Moore, E. Nye, B. Spencer-Dene and J. Bee for technical support with mice and mouse tissue. We also thank the Flow Cytometry Unit, the Bioinformatics & Biostatistics Unit and the *In vivo* Imaging Facility for technical assistance. We are grateful to Cancer Research UK for the funding that has allowed this work.

**Author Contributions** S.K.W. organised and performed all experiments, helped design experiments, interpreted data and helped with manuscript preparation. I.M. conceived and supervised the study, designed experiments, interpreted the data, assisted with some aspects of the experiments and wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.M. (ilaria.malanchi@crick.ac.uk).



## METHODS

**Mouse strains.** The MMTV-PyMT mice were a gift from E. Sahai, MMTV-PyMT actin-GFP (mice expressing green fluorescent protein under the control of the actin promoter), *Gcsf*-null and *Rag1*-null mice were a gift from J. Huelsken, MMTV-PyMT actin-luciferase (mice expressing firefly luciferase under the control of the actin promoter) transgenic line was a gift from D. Bonnet, Rosa26R-eGFP-DTA mice were a gift from C. Reis e Sousa. Ela2-Cre knock-in mice and *Alox5*-null mice were purchased from European Mouse Mutant Archive (EMMA) and Jackson Laboratory, respectively. All mouse strains have been described previously<sup>30–37</sup>. All strains of mice were in >10 generations FVB/N and/or C57BL/6 background except *Gcsf*-null, Ela2-Cre and Rosa26R-eGFP-DTA mice that were used in mixed background with littermate controls. Female mice were used between 6–9 weeks of age, except spontaneous cancer models. Breeding and all animal procedures were performed in accordance with UK Home Office regulations under project license PPL/80/2531.

**Mouse experiments.** Where applicable, mice were anaesthetized with IsoFlo (isoflurane, Abbott Animal Health) and temporally treated with the analgesics Vetgesic (Alstoe Animal Health) and/or Rimadyl (Pfizer Animal Health). For tumour studies under the project licence PPL/80/2531, the overruling determinant was animal welfare. The National Cancer Research Institute (NCRI) Guidelines for the Welfare and Use of Animals in Cancer Research were followed. When assessing primary tumour growth, a mean diameter of 1.5 cm for single tumours was not exceeded. However, for multifocal disease such as MMTV-PyMT cancer, provided that there were no additional adverse welfare consequences for the animal, the total superficial tumour burden was allowed to exceed these dimensions when essential for the achievement of the scientific objective, namely spontaneous metastasis. Mice were monitored daily for signs of adverse effects. The source data for primary tumour growth are in Supplementary Fig. 3.

**Tumour cell transplantations and induction of experimental metastasis.** FVB/N wild-type mice were used for MMTV-PyMT tumour cell transplantations to isolate lung neutrophils. *Rag1*-null mice were used when using human or mouse GFP or luciferase-expressing tumour cells. Primary MMTV-PyMT, MMTV-PyMT actin-GFP or MMTV-PyMT actin-luciferase cells ( $10^5$ – $10^6$  cells per injection), the unmarked or stably mouse phosphoglycerate kinase 1 (PGK) promoter-GFP-expressing mouse mammary cancer cell line 4T1 ( $10^5$  cells per injection) and the unmarked or stably actin-GFP-expressing human breast cancer cell line MDA-MB-231 ( $1$ – $2 \times 10^6$  cells per injection) were used. For experimental metastasis, tumour cells were re-suspended in 100  $\mu$ l PBS and tail vein injected. For orthotopic transplantations, tumour cells were re-suspended in 50  $\mu$ l growth-factor-reduced Matrigel (Costar) and transplanted within the fourth mammary fat pad on both flanks (MMTV-PyMT and MDA-MB-231 cells) or one flank only (4T1 cells).

**Neutrophilia and lung immune cell infiltration in MMTV-PyMT<sup>+</sup> mice.** MMTV-PyMT<sup>+</sup> mice that spontaneously developed a primary tumour and had visible lung metastasis were used to determine immune cell presence in the lung and neutrophil presence in other organs together with tumour-free littermate controls. For determination of timing and dynamics of lung infiltration by neutrophils and cancer cells, MMTV-PyMT<sup>+</sup> mice harbouring 1.5–2 g spontaneously developed tumours were used. Neutrophil infiltration was quantified by flow cytometry and histological staining of lung sections for S100A9 and cancer cell presence by examination of six histological lung sections (100  $\mu$ m apart) for PyMT staining to confirm the pre-metastatic state. The timing of neutrophil infiltration into the pre-metastatic lung before cancer cells was confirmed in FVB/N wild-type mice carrying two primary tumours originating from orthotopic injection of primary MMTV-PyMT cancer cells and used for analysis (daily treated with anti-Ly6G or control IgG antibody starting 24 h before tumour cell implantation).

**Analysis of MMTV-PyMT<sup>+</sup> G-CSF and MMTV-PyMT<sup>+</sup> Ela2-DTA mice.** Mice were culled and analysed about 6 weeks after spontaneous primary tumour onset; no differences were observed in tumour onset among the different genotypes.

**Treatments with neutrophil-blocking antibody anti-Ly6G or Zil.** Rat anti-Ly6G antibody<sup>38,39</sup> (12.5  $\mu$ g per mouse; clone 1A8 from BioXcell) or rat IgG isotype control (provided by the Cell Services Unit of The Crick Institute) in 100  $\mu$ l saline were administered daily via intraperitoneal injection. Zil (LKT Laboratories) dissolved in DMSO (Sigma) or DMSO alone was fed to mice by pipetting on the back of the tongue once a day at a dosage of 100  $\mu$ g Zil per g mouse weight.

**Lung colonization by cancer cells after neutrophil depletion or Zil treatment.** *Rag1*-null mice were orthotopically transplanted with unlabelled mammary tumour cells 4 weeks before labelled tumour cell injection via the tail vein (MMTV-PyMT and 4T1  $10^5$  cells, MDA-MB-231  $10^6$  cells). Anti-Ly6G or Zil treatment for 2 weeks (except 4T1, 10 days) started 1 day before intravenous injection of cancer cells. Then, total primary tumour burden, neutrophil presence in the lung, spontaneous lung metastasis incidence from the transplanted primary tumour and/or

experimentally induced lung metastasis originating from the intravenously injected cancer cells was analysed.

Of note, exclusively experimental metastasis are present in lung harbouring MDA-MB-231 cells, while predominantly spontaneous metastases are visible in lung harbouring 4T1 cells due to the high spontaneous metastasis rate of primary 4T1 tumours. Only GFP<sup>+</sup> experimental metastasis induced by cancer cell injection was quantified in these experiments.

**Tumour/metastasis initiation potential assay *in vivo*.** Primary MMTV-PyMT cells were either cell sorted for BLT2 and/or CysLT2 presence or absence, or treated for 3 days on collagen-coated dishes with either neutrophil-conditioned medium or LTB4 and LTC/D/E4. Subsequently,  $10^3$ – $10^4$  cells were orthotopically transplanted into the mammary gland or  $10^6$  cells injected via the tail vein into *Rag1*-null mice and mammary tumour growth or lung metastasis incidence analysed about 3 weeks thereafter.

**MICs or metastasis quantification after neutrophil/LuN injection.** To analyse total cancer cells at early stages, *Rag1*-null mice were injected with  $0.5$ – $1 \times 10^6$  MMTV-PyMT actin-GFP cells via the tail vein followed 12 h later by intravenous injection of  $25 \times 10^6$  neutrophils (freshly isolated from MMTV-PyMT tumour-transplanted mice) or 12, 24 and 36 h later by intravenous injection of 200  $\mu$ l lung neutrophil-conditioned or control sphere medium (described later). Cancer cells in the lung were analysed 3 days after the initial tumour cell injection for frequencies of CD90<sup>+</sup> MICs among GFP<sup>+</sup>CD24<sup>+</sup> (non-MIC) cancer cells. For determination of effects of neutrophils or neutrophil-conditioned medium on metastatic burden, *Rag1*-null mice were intravenously injected with  $1$ – $10 \times 10^5$  MMTV-PyMT actin-GFP or actin-luciferase cells followed immediately, 2 and 4 days later, by injection of  $25 \times 10^6$  neutrophils or 3–5 times every 12 h by injection of 200  $\mu$ l lung neutrophil-conditioned medium. Metastatic burden was determined by flow cytometric analysis of GFP<sup>+</sup> cancer cells 1 week or bioluminescence imaging of luciferase<sup>+</sup> cancer cells 2–4 weeks thereafter, respectively.

**Analysis of functional effects of G-CSF deficiency in MMTV-PyMT cancer cells.** *Rag1*-null mice were transplanted with  $10^6$  *Gcsf*-null primary MMTV-PyMT cancer cells into two mammary glands and tumour growth, spontaneous metastatic incidence and neutrophil presence in the lung were analysed 4 weeks thereafter.

**Bone marrow transplantation and semi-quantitative PCR.** C57BL/6 wild-type mice were lethally irradiated (dosage:  $2 \times 600$  rad, 4 h apart) and 24 h later injected via the tail vein with  $2 \times 10^6$  bone marrow cells freshly isolated from C57BL/6 or *Alox5*-null donor mice. Bone marrow chimaeric mice were orthotopically transplanted with  $10^6$  MMTV-PyMT cells into the fourth mammary fat pad on both sides 8 weeks after bone marrow reconstitution and primary tumour size, neutrophil infiltration into the lung and lung metastasis were analysed 6 weeks later. Chimaeric mice were generated in a pure C57BL/6 background, therefore MMTV-PyMT cells from the same background were used to generate primary tumours. In this lower tumorigenic background, metastasis only occurs in about 50% of the mice. No alteration in this penetrance was observed between wild-type and *Alox5*-null bone marrow chimaeric mice. Figure 4b quantifies animals harbouring metastatic disease.

Percentage of bone marrow reconstitution was calculated by isolating total DNA from bone marrow of chimaeras and semi-quantitative PCR with a calibration curve from 100% wild-type DNA mixed at defined ratios with 100% *Alox5*-null DNA. PCR was performed using Redtag reagents (Sigma) (primers are listed in Supplementary Information) and 25 amplification cycles before loading an agarose gel. Ratio between wild-type and *Alox5*-null band was calculated for every mouse and percentage chimaerism was determined by comparison with calibration curve. Chimaerism was consistently between 80 and 96%.

**Tumour and metastasis burden evaluation.** See Supplementary Methods.

***In vivo* luciferase-activity detection.** Mice inoculated with actin-luciferase-expressing MMTV-PyMT cells were shaved around the chest area and injected with 3 mg Xenolight D-luciferin potassium salt (PerkinElmer) in PBS into the peritoneum 5 min before imaging for at least 45 min using the IVIS Spectrum Pre-clinical *In vivo* Imaging System (PerkinElmer). The maximum bioluminescence intensity signal for the lung of every mouse was determined using Living Image 4.3.1 software.

**Tissue staining, immunohistochemistry and light microscopy.** Mouse lung tissue was fixed in 4% paraformaldehyde in PBS for 24 h and embedded in paraffin blocks. Four-micrometre sections were stained. The breast cancer tissue array paired with metastatic tumours, 96 samples (1.5 mm), was purchased from Abcam (ab178118). H&E staining was performed according to standard procedures.

For immunohistochemistry, either secondary horseradish peroxidase (HRP)-conjugated antibodies were used in combination with DAP Peroxidase substrate or the VECTASTAIN ABC kit (all Vector Laboratories) according to the manufacturer's instructions. Specific primary antibodies were used (see Supplementary

Information), visualization of cell nuclei was performed with haematoxylin and analysis employed the Nikon Eclipse 90i light microscope and NIS-elements software.

**Scoring of LTR expression in breast cancer tissue and lymph node metastasis.** See Supplementary Methods.

**Tissue digestion for cell isolation or analysis.** MMTV-PyMT cell isolation was described in detail previously<sup>8</sup>. In brief, primary MMTV-PyMT tumours, liver, spleen and lung were dissected, minced, and digested with Liberase (Roche) and DNaseI (Sigma) in HBSS and passed through a 100 µm cell strainer. Some tumour cells were used for cell culture at this point. Bone marrow cells were isolated by crushing the femur and tibia and blood collected via bleeding from the tail vein with heparin (Sigma) as a coagulant. For flow cytometric analysis or further purification, single-cell suspensions of tumour, liver, spleen, lung, bone marrow and blood were subjected to hypotonic lysis (Red Blood Cell Lysis Solution, Miltenyi) to remove erythrocytes and washed with 1 × PBS/2 mM EDTA/0.5% BSA.

**Flow cytometry and cell sorting.** Prepared single-cell suspensions of mouse tissues and *in vitro* treated cancer cells were incubated with mouse FcR Blocking Reagent (Miltenyi) followed by incubation with (a combination) of specific pre-labelled antibodies or in combination with fluorescently labelled secondary antibodies (Invitrogen) (see Supplementary Information). Dead cells were stained with 4',6-diamidino-2-phenylindole (DAPI) or propidium iodide (PI; both Sigma). The LSRFortessa cell analyser running FACSDiva software (BD Biosciences) and FlowJo software was used. Tumour cells were flow-sorted using the Influx cell sorter running FACS Sortware sorter software (BD Biosciences). MMTV-PyMT cells were used in experiments immediately after sorting and sorted 4T1 cells cultured in adherent conditions for 3 days before western blot analysis.

**Neutrophil isolation and neutrophil-conditioned medium.** Freshly isolated lung cells from wild-type mice orthotopically transplanted with MMTV-PyMT tumours were incubated with mouse FcR Blocking Reagent (Miltenyi), APC-coupled anti-Ly6G (clone 1A8) antibody (BD Bioscience) followed by incubation with magnetic anti-APC microbeads (Miltenyi). Magnetically labelled neutrophils were isolated using LS columns (Miltenyi) according to the manufacturer's instructions. Neutrophil purity and viability was measured by flow cytometry. Some isolated Ly6G<sup>+</sup> cells were smeared onto a glass slide and air-dried overnight followed by H&E staining to evaluate cell morphology. Remaining neutrophils were kept in sphere medium at a concentration of 10<sup>6</sup> neutrophils per 150 µl medium for 14 h to allow conditioning. Neutrophils and cell debris were removed by centrifugation and conditioned medium occasionally snap-frozen before use.

**Cell culture and *in vitro* cancer cell treatments.** All used cell lines were provided by the Cell Services Unit of The Crick Institute, which routinely tests for *Mycoplasma* contamination and were not further authenticated in our laboratory. Cell lines were cultured in DMEM medium supplemented with 10% fetal bovine serum (DMEM/FCS, both Invitrogen). Freshly isolated MMTV-PyMT cells were cultured overnight on PureCol collagen (Advanced Biomatrix)-coated dishes in growth medium DMEM/F12 with 2% FBS, 20 ng ml<sup>-1</sup> EGF (Invitrogen) and 10 µg ml<sup>-1</sup> insulin (Sigma) before use in experiments. All *in vitro* and *in vivo* experiments involving primary MMTV-PyMT cells were performed with at least two primary tumour cell preparations from different spontaneous MMTV-PyMT<sup>+</sup> mice. Unless otherwise specified, each *in vitro* and *in vivo* experiment was performed with a different tumour cell preparation.

Primary MMTV-PyMT cells were cultured in sphere medium on collagen-coated dishes, 4T1 and MDA-MB-231 cells in DMEM/FCS on uncoated dishes or in non-attachment conditions for the indicated periods of time under presence of (as indicated for every experiment): control sphere medium, neutrophil-conditioned medium, 100% ethanol control (EtOH, Sigma), DMSO control, 1 µM LTB4, 100 nM LTC/D/E4 (Cysteinyl Leukotriene HPLC Mixture 1), 3 µM BLT2 inhibitor LY255283, 0.3 µM CysLT2 inhibitor BAY-u9773 (all Cayman Chemical), 1 µM Zil and/or 1 nM pan-MEK inhibitor PD0325901 (provided by J. Downwards) followed by further tests or analysis.

**Sphere formation assay.** The sphere formation assay was described previously<sup>8</sup>. In brief, 10<sup>4</sup> total MMTV-PyMT or flow-sorted cells per well were plated in ultra-low-attachment 96-well plates (Costar) in 100 µl sphere medium DMEM/F12 supplemented with B-27, 20 ng ml<sup>-1</sup> EGF, 20 ng ml<sup>-1</sup> FGF (all Invitrogen) and 4 µg ml<sup>-1</sup> heparin (Sigma) or neutrophil-conditioned medium. After 7–10 days, if not otherwise indicated, all formed spheres were quantified from images taken with the inverted Leica DM IRBE light and fluorescence microscope. The area of the plane passing through the sphere centre was measured for every sphere (sphere size) using ImageJ software and the areas of all formed spheres were summed up. The obtained number was divided by total number of plated cells. This value represents the sphere formation index (SFI) per cell for every experimental group. Freshly isolated MMTV-PyMT cells were either only treated for 3 days in adherent conditions before sphere assay or directly treated during the sphere assay with

neutrophil-conditioned medium or LTB4 and/or LTC/D/E4 or Zil, as indicated. When cells were passaged, cells were quantified by cell counting and re-plated in equal numbers per well for the next passage approximately every 7 to 10 days.

***In vitro* and *in vivo* BrdU incorporation assay.** Rag1-null mice carrying MMTV-PyMT tumours were treated daily for 3 days with Zil and intravenously injected with 10<sup>5</sup> GFP-expressing MMTV-PyMT cancer cells. BrdU (1 mg per mouse) was intraperitoneally injected 18 h after GFP<sup>+</sup> cancer cells and lungs were harvested and digested 6 h later. *In vitro* 3-day MMTV-PyMT or 4T1 cells treated as indicated in adherent conditions were pulsed with 30 µM BrdU (Sigma) for 3 h and harvested. Cells were incubated with fluorescently labelled anti-CD24 and/or anti-CD90.1 antibody if indicated. BrdU Flow Kit (BD Bioscience) was used for staining followed by analysis by flow cytometry.

***In vitro* quantification of primary MICs and sub-pools of cancer cell lines.**

Primary MMTV-PyMT cells were cultured on collagen-coated dishes for 3 days supplemented with either LTB4 and LTC/D/E4 or Zil followed by incubation with fluorescently labelled anti-CD90.1 and anti-CD24 antibodies and analysis by flow cytometry. 4T1 and MDA-MB-231 cell lines were cultured in DMEM/FCS supplemented with LTB4 and LTC/D/E4 for 3 days in adherent conditions followed by either staining with fluorescently labelled anti-CD49f, anti-BLT2, anti-CysLT2 and/or anti-CD44 antibodies or using the ALDEFLUOR kit (StemCell Technologies) according to the manufacturer's instructions and analysed by flow cytometry.

**RNA expression/quantitative real-time PCR.** Neutrophils were freshly isolated from the lungs of wild-type or MMTV-PyMT tumour-bearing mice. RNA isolation was performed using MagMAX-96 Total RNA Isolation Kit and cDNA synthesis using SuperScript III Reverse Transcriptase. Quantitative PCR reactions were performed using EXPRESS SYBR GreenER reagents with the Applied Biosystems 7500 Fast Real-Time PCR System (all Invitrogen) and specific primers (see Supplementary Information).

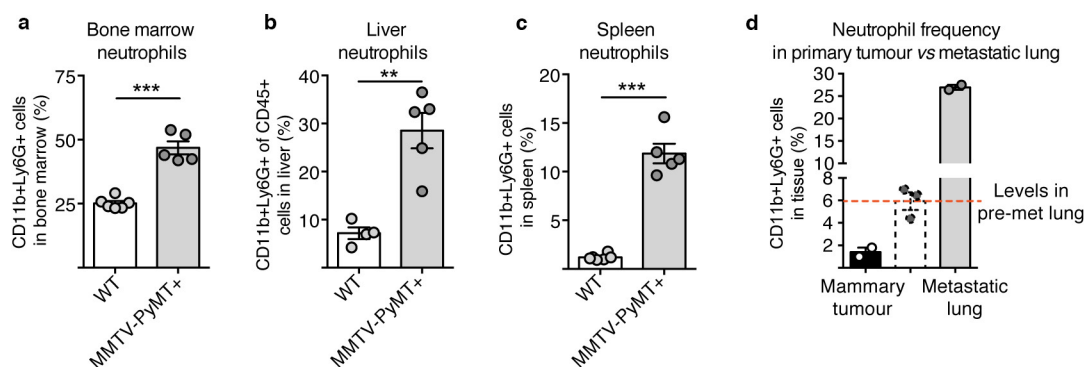
**Enzyme immunoassay and parameter enzyme-linked immunosorbent assay.** Ethanol was used to precipitate protein from cell culture medium before analysis using either the enzyme immunoassays (EIAs) LTC/D/E4 Biotrak EIA System (Amersham) or the LTB4 EIA Kit (Cayman Chemical) according to the manufacturer's instructions.

**Western blot analysis and protein detection.** Primary MMTV-PyMT cells grown on collagen-coated dishes were cultured overnight in DMEM/F12 with B-27, and 4 µg ml<sup>-1</sup> heparin (Sigma) before treatment with 1 µM LTB4 or 100 nM LTC/D/E4. Unsorted or sorted LTR-reduced 4T1 cells were stimulated with LTB4, LTC/D/E4, BLT2 inhibitor LY255283 and/or CysLT2 inhibitor BAY-u9773 as indicated. Cells were washed and protein isolated using RIPA buffer (25 mM Tris-hydrogen chloride pH 7.6, 50 mM sodium chloride, 1% NP-40, 1% sodium deoxycholate, 0.1% sodium dodecyl sulfate) freshly supplemented with 1 µM sodium pyrophosphate, 1 µM B-glycerophosphate, 1 µM sodium vanadomoxide, 1 µM sodium fluoride, 1 µM sodium molybdate (all Sigma) and cOmplete ULTRA Tablets (Roche), and processed by standard western blot techniques. Membranes were blocked with 5%BSA in PBS with 0.5% Tween-20 (Sigma) and incubated with specific primary antibodies (see Supplementary Information). ECL Western Blotting System including secondary antibodies and Hyperfilm ECL (both Amersham) were used. Protein lysates of 3 h LTB4-stimulated MDA-MB-231 cells were analysed using the Proteome Profiler Human Phospho-Kinase Array Kit (R&D systems) according to the manufacturer's instructions. Western blot quantification was performed on scanned films using ImageJ software.

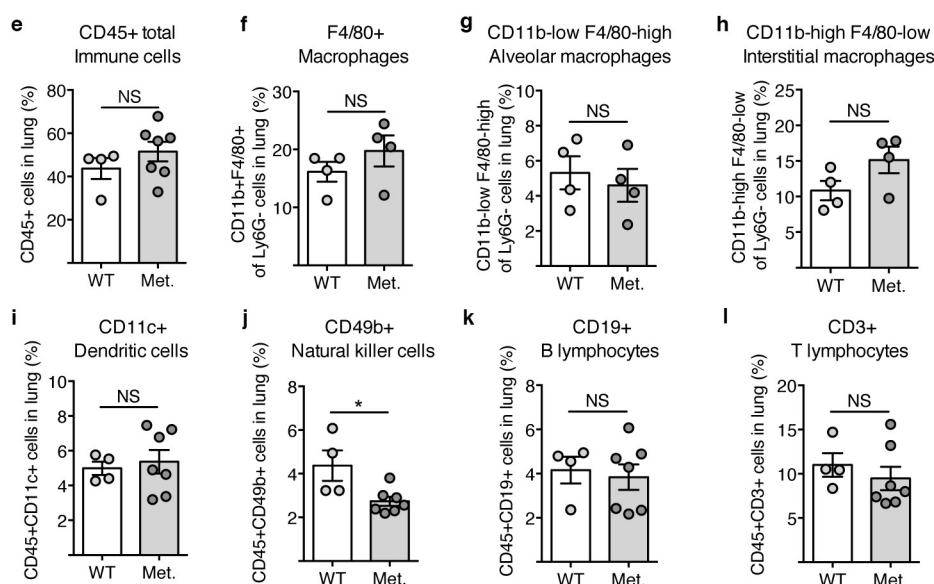
**Statistical analysis.** Data analyses used GraphPad Prism version 7. The data are presented as mean ± standard error of the mean, individual values, 'scatter plot with Tukey box and whiskers' and/or 'scatter plot with column bar' graphs and were analysed using Student's *t*-tests (paired or unpaired according to the experimental setting), Mann–Whitney tests, one-sample *t*-tests and two-way ANOVA as indicated in the legends. Data were pooled from at least two experiments, except Fig. 4c, i, k and Extended Data Figs 2d, 4d–m, 5a, b, f, h, 6k, 10e, in which data are at least biological triplicates generated in parallel. Two-way ANOVA was performed when the control groups between experiments were significantly different. Western blot in Extended Data Fig. 8i, k, the proteome profiler dot blot in Extended Data Fig. 8d and BrdU incorporation of 4T1 cells in Extended Data Fig. 10k were performed once. Extended Data Fig. 3b (mRNA expression) compares biological triplicates of the pre-metastatic to a representative control (wild-type) value. The experiments were not randomized and there was no blinding as animals or samples were marked. No statistical methods were used to predetermine sample sizes. Sample sizes were based on previous experience with the models<sup>8,14</sup>, *n* values represent biological replicates, with the exception of the sphere assays, for which both technical and biological replicates are shown.

Differences were considered significant when *P* < 0.05 and are indicated as NS, not significant, \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001.

30. Guy, C. T., Cardiff, R. D. & Muller, W. J. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. *Mol. Cell. Biol.* **12**, 954–961 (1992).
31. Okabe, M., Ikawa, M., Kominami, K., Nakanishi, T. & Nishimune, Y. 'Green mice' as a source of ubiquitous green cells. *FEBS Lett.* **407**, 313–319 (1997).
32. Lieschke, G. J. *et al.* Mice lacking granulocyte colony-stimulating factor have chronic neutropenia, granulocyte and macrophage progenitor cell deficiency, and impaired neutrophil mobilization. *Blood* **84**, 1737–1746 (1994).
33. Mombaerts, P. *et al.* RAG-1-deficient mice have no mature B and T lymphocytes. *Cell* **68**, 869–877 (1992).
34. Cao, Y. A. *et al.* Shifting foci of hematopoiesis during reconstitution from single stem cells. *Proc. Natl Acad. Sci. USA* **101**, 221–226 (2004).
35. Ivanova, A. *et al.* *In vivo* genetic ablation by Cre-mediated expression of diphtheria toxin fragment A. *Genesis* **43**, 129–135 (2005).
36. Tkalec, J. *et al.* Impaired immunity and enhanced resistance to endotoxin in the absence of neutrophil elastase and cathepsin G. *Immunity* **12**, 201–210 (2000).
37. Chen, X. S., Sheller, J. R., Johnson, E. N. & Funk, C. D. Role of leukotrienes revealed by targeted disruption of the 5-lipoxygenase gene. *Nature* **372**, 179–182 (1994).
38. Daley, J. M., Thomay, A. A., Connolly, M. D., Reichner, J. S. & Albina, J. E. Use of Ly6G-specific monoclonal antibody to deplete neutrophils in mice. *J. Leukoc. Biol.* **83**, 64–70 (2008).
39. Bao, Y. & Cao, X. Revisiting the protective and pathogenic roles of neutrophils: Ly-6G is key! *Eur. J. Immunol.* **41**, 2535–2538 (2011).

Systemic increase of neutrophils in MMTV-PyMT<sup>+</sup> mice:

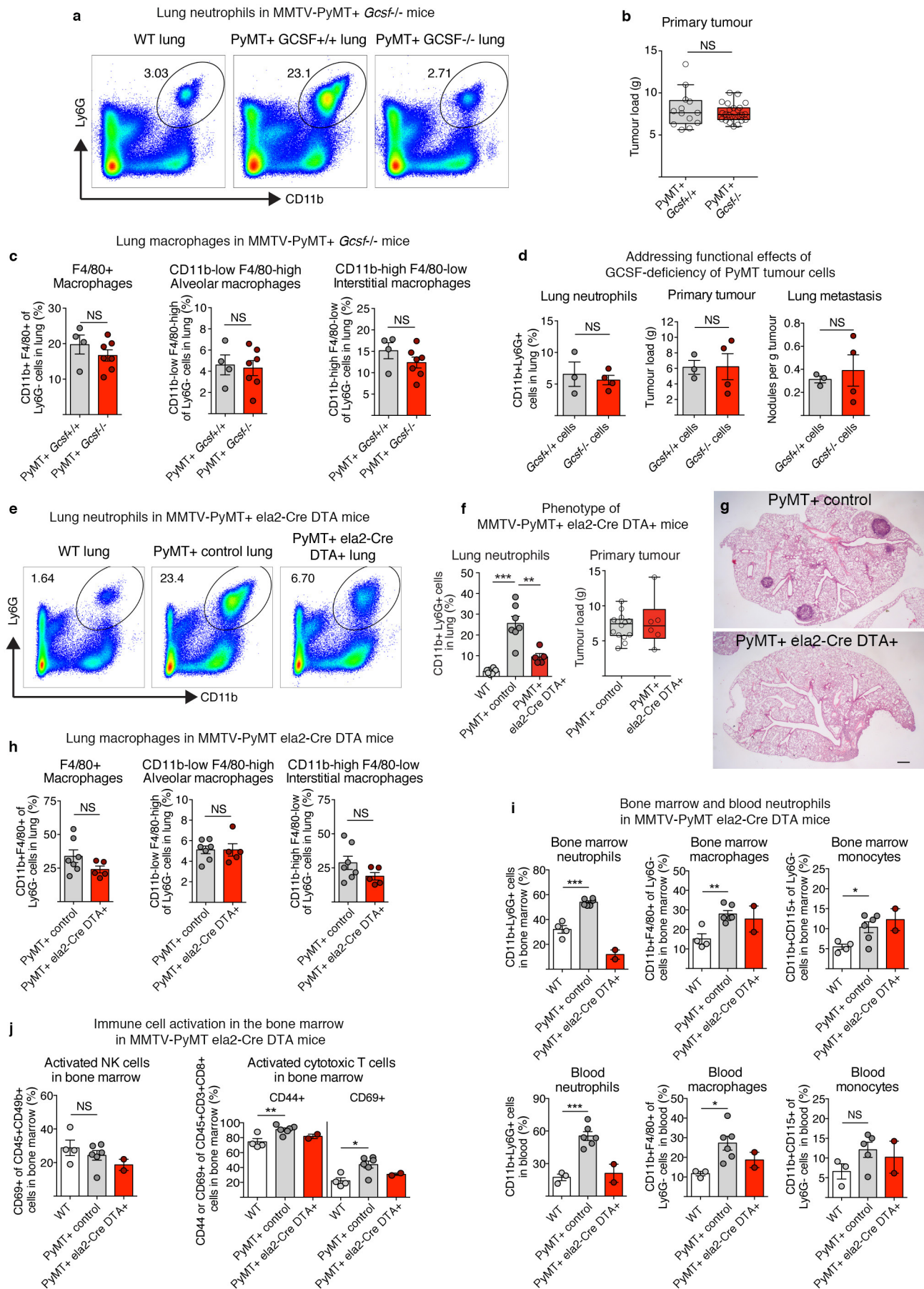
## Characterisation of immune cell presence in the metastatic lung



**Extended Data Figure 1 | Mammary tumour-bearing MMTV-PyMT<sup>+</sup> mice show specifically neutrophilia in the metastatic lung.** a–c, Flow cytometric quantification of CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophils in the bone marrow ( $n = 6$  (wild type),  $n = 5$  (MMTV-PyMT<sup>+</sup>)) (a), liver ( $n = 4$  (wild type),  $n = 5$  (MMTV-PyMT<sup>+</sup>)) (b) and spleen ( $n = 6$  (wild type),  $n = 5$  (MMTV-PyMT<sup>+</sup>)) (c) of wild-type (WT) and tumour-bearing MMTV-PyMT<sup>+</sup> mice. d, Quantification of neutrophils in the tumour and metastatic lung of MMTV-PyMT<sup>+</sup> mice ( $n = 2$  per group), pre-metastatic lung neutrophil levels depicted in Fig. 1a are shown for comparison in dashed lines. Met., metastatic. e–l, Flow cytometric quantification of immune cell frequencies in wild-type and metastatic

lungs of MMTV-PyMT<sup>+</sup> mice ( $n = 4$  (wild type),  $n = 7$  (metastatic) if not otherwise indicated) including CD45<sup>+</sup> total immune cells (e), total CD11b<sup>+</sup>F4/80<sup>+</sup> macrophages (f) ( $n = 4$  (wild type),  $n = 4$  (metastatic)), the CD11b<sup>low</sup>F4/80<sup>high</sup> alveolar macrophage subpopulation ( $n = 4$  (wild type),  $n = 4$  (metastatic)) (g), the CD11b<sup>high</sup>F4/80<sup>low</sup> interstitial macrophage subpopulation ( $n = 4$ /WT,  $n = 4$ /Met.) (h), CD45<sup>+</sup>CD11c<sup>+</sup> dendritic cells (i), CD45<sup>+</sup>CD49b<sup>+</sup> NK cells (j), CD45<sup>+</sup>CD19<sup>+</sup> B lymphocytes (k) and CD45<sup>+</sup>CD3<sup>+</sup> T lymphocytes (l). Statistical analysis by two-sided *t*-test. Data are represented as mean  $\pm$  s.e.m. NS, not significant, \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

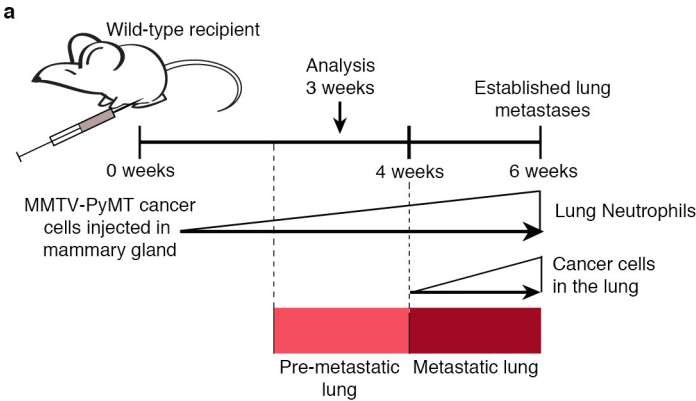




Extended Data Figure 2 | See next page for caption.

**Extended Data Figure 2 | Analysis of MMTV-PyMT<sup>+</sup>Gcsf<sup>-/-</sup> mice, G-CSF-deficient MMTV-PyMT cancer cells and MMTV-PyMT<sup>+</sup>Ela2-Cre-DTA<sup>+</sup> mice.** **a**, Representative flow cytometric analysis of CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophils in the lung of wild-type and tumour-bearing MMTV-PyMT<sup>+</sup>Gcsf<sup>+/+</sup> and MMTV-PyMT<sup>+</sup>Gcsf<sup>-/-</sup> mice. **b**, Primary mammary tumour burden of MMTV-PyMT<sup>+</sup>Gcsf<sup>+/+</sup> ( $n = 13$ ) or MMTV-PyMT<sup>+</sup>Gcsf<sup>-/-</sup> ( $n = 24$ ) mice. **c**, Flow cytometric quantification of frequencies of total CD11b<sup>+</sup>F4/80<sup>+</sup> macrophages (left), the CD11b<sup>low</sup>F4/80<sup>high</sup> alveolar macrophage subpopulation (middle) and the CD11b<sup>high</sup>F4/80<sup>low</sup> interstitial macrophage subpopulation (right) in the lung of tumour-bearing MMTV-PyMT<sup>+</sup>Gcsf<sup>+/+</sup> ( $n = 4$ ) and MMTV-PyMT<sup>+</sup>Gcsf<sup>-/-</sup> ( $n = 7$ ) mice. **d**, MMTV-PyMT<sup>+</sup>Gcsf<sup>-/-</sup> primary cancer cells were freshly isolated and grafted onto two mammary glands of Rag1-null mice ( $10^6$  cells per injection) and analysed 5 weeks thereafter. CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophil presence in the lung was assessed by flow cytometry (left), primary tumour burden was assessed by weighing (middle) and spontaneous lung metastasis incidence was assessed by quantification of visible surface lung metastases relative to tumour load (right) ( $n = 3$  (Gcsf<sup>+/+</sup>),  $n = 4$  (Gcsf<sup>-/-</sup>)). **e–g**, Analysis of tumour-bearing MMTV-PyMT<sup>+</sup> control and MMTV-PyMT<sup>+</sup>Ela2-Cre-DTA<sup>+</sup> mice. Representative flow cytometric analysis of CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophils in the lung (**e**). Lung neutrophil quantification ( $n = 8$  (wild type),  $n = 7$  (PyMT+ control),  $n = 5$  (PyMT+Ela2-Cre-DTA+)) (**f**, left) and primary mammary tumour burden ( $n = 14$  (PyMT+ control),  $n = 6$

(PyMT+Ela2-Cre-DTA+)) (**f**, right) with representative H&E-stained histological lung sections (**g**). Scale bar, 500  $\mu$ m. **h**, Flow cytometric quantification of frequencies of total CD11b<sup>+</sup>F4/80<sup>+</sup> macrophages (left), the CD11b<sup>low</sup>F4/80<sup>high</sup> alveolar macrophage subpopulation (middle) and the CD11b<sup>high</sup>F4/80<sup>low</sup> interstitial macrophage subpopulation (right) in the lung of tumour-bearing MMTV-PyMT<sup>+</sup> control ( $n = 7$ ) and MMTV-PyMT<sup>+</sup>Ela2-Cre-DTA<sup>+</sup> ( $n = 5$ ) mice. **i**, Frequencies of bone marrow (top) and blood (bottom) CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophils (left; blood  $n = 3$  (wild type),  $n = 6$  (PyMT+ control)), CD11b<sup>+</sup>F4/80<sup>+</sup> macrophages (middle; blood  $n = 3$  (wild type),  $n = 6$  (PyMT+ control)) and CD11b<sup>+</sup>CD115<sup>+</sup> monocytes (right; blood  $n = 3$  (wild type),  $n = 5$  (PyMT+ control)) in wild-type, MMTV-PyMT<sup>+</sup> control and MMTV-PyMT<sup>+</sup>Ela2-Cre-DTA<sup>+</sup> mice analysed by flow cytometry ( $n = 4$  (wild type),  $n = 6$  (PyMT+ control),  $n = 2$  (PyMT+Ela2-Cre-DTA+)) if not otherwise indicated). **j**, Exclusion of immune responses against DTA expression in the bone marrow by analysis of NK-cell (left) and cytotoxic T-cell (right) activation. Flow cytometric quantification of activated CD69<sup>+</sup> among total CD45<sup>+</sup>CD49b<sup>+</sup> NK cells as well as activated CD44<sup>+</sup> or CD69<sup>+</sup> among total CD45<sup>+</sup>CD3<sup>+</sup>CD8<sup>+</sup> cytotoxic T cells in the bone marrow of wild-type ( $n = 4$ ), MMTV-PyMT<sup>+</sup> control ( $n = 6$ ) and MMTV-PyMT<sup>+</sup>Ela2-Cre-DTA<sup>+</sup> ( $n = 2$ ) mice. Statistical analysis by two-sided *t*-test. Data are represented as mean  $\pm$  s.e.m. NS, not significant, \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .



**b** Characterisation of neutrophils accumulating in the pre-metastatic lung

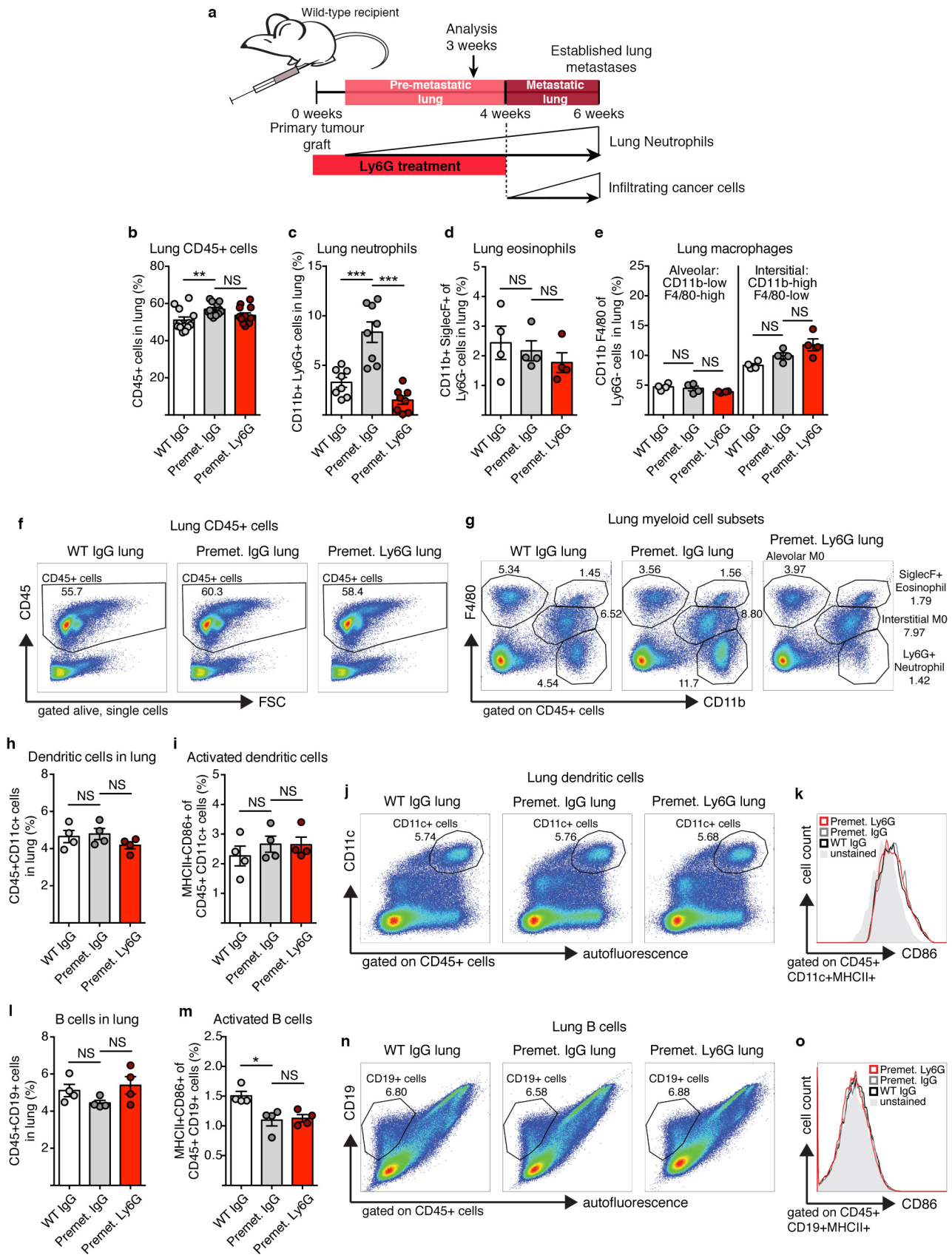
Analysis of Neutrophils in the lung				
	Wild-type	Pre-metastatic		
Cytometric analysis:	Mean intensity of CD11b+ Ly6G+ cells		Significance	
Cell size (FSC)	84295 ± 199.0	87223 ± 628.1	N=4, P=0.0044	*
Cell granularity (SSC)	54932 ± 1467	62774 ± 2020	N=4, P=0.0200	**
Surface expression:	Percent positive of CD11b+ Ly6G+ cells			
CXCR2+	97.90 ± 0.524	98.45 ± 0.1500	N=4, P=0.3522	
CD31+	31.58 ± 3.728	49.45 ± 1.900	N=4, P=0.0053	**
MHC-I+	30.60 ± 1.696	28.77 ± 2.381	N=4, P=0.5533	
MHC-II+	28.18 ± 0.820	22.82 ± 1.415	N=4, P=0.0170	*
ICAM1+	23.90 ± 2.515	23.90 ± 2.515	N=3/WT, N=6/Pre-met, P=0.9008	
Fas+	99.35 ± 0.050	99.35 ± 0.050	N=2/WT, N=3/Pre-met, P=0.1697	
mRNA expression:	Fold-change relative to wildtype		Neutrophil purity ≥ 90%	
<i>TNFi</i>	1	0.9585 ± 0.1735	P=0.8334	
<i>Arginase 1</i>	1	0.7667 ± 0.1924	P=0.3492	
<i>VEGF-A</i>	1	0.6814 ± 0.1494	P=0.1666	
<i>CCL2</i>	1	0.4157 ± 0.0932	P=0.0245	*
<i>CCL3</i>	1	0.7391 ± 0.1584	P=0.2414	
<i>iNOS</i>	1	1.0360 ± 0.4653	P=0.9506	
<i>CCL5</i>	1	0.0517 ± 0.0042	P<0.0001	***

**Extended Data Figure 3 | Comparison of wild-type lung neutrophils with tumour-induced, pre-metastatic lung neutrophils.**

**a**, Representation of timing and dynamics of neutrophil and cancer cell infiltration into the lung of mice grafted with two mammary tumours by orthotopic injection of  $10^6$  MMTV-PyMT tumour cells. **b**, Flow cytometric analysis for cell size (forward scatter (FSC)), granularity (side scatter (SSC)) and expression of surface markers CXCR2, CD31, MHC-I, MHC-II, ICAM1 and Fas ( $n$  is indicated) as well as mRNA

expression analysis of *Tnfa*, arginase 1, *Vegfa*, *Ccl2*, *Ccl3*, *iNOS* (also known as *Nos2*) and *Ccl5* by quantitative polymerase chain reaction (PCR) of CD11b<sup>+</sup>Ly6G<sup>+</sup> wild-type (WT) or pre-metastatic (Pre-met.) lung neutrophils 3 weeks after primary tumour graft ( $n = 3$  (pre-metastatic compared with one normal lung reference)). Statistical analysis by two-sided  $t$ -test (flow cytometry) and one-sample  $t$ -test (mRNA). Data are represented as mean  $\pm$  s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

Addressing the immunologic response in tumour-bearing mice:



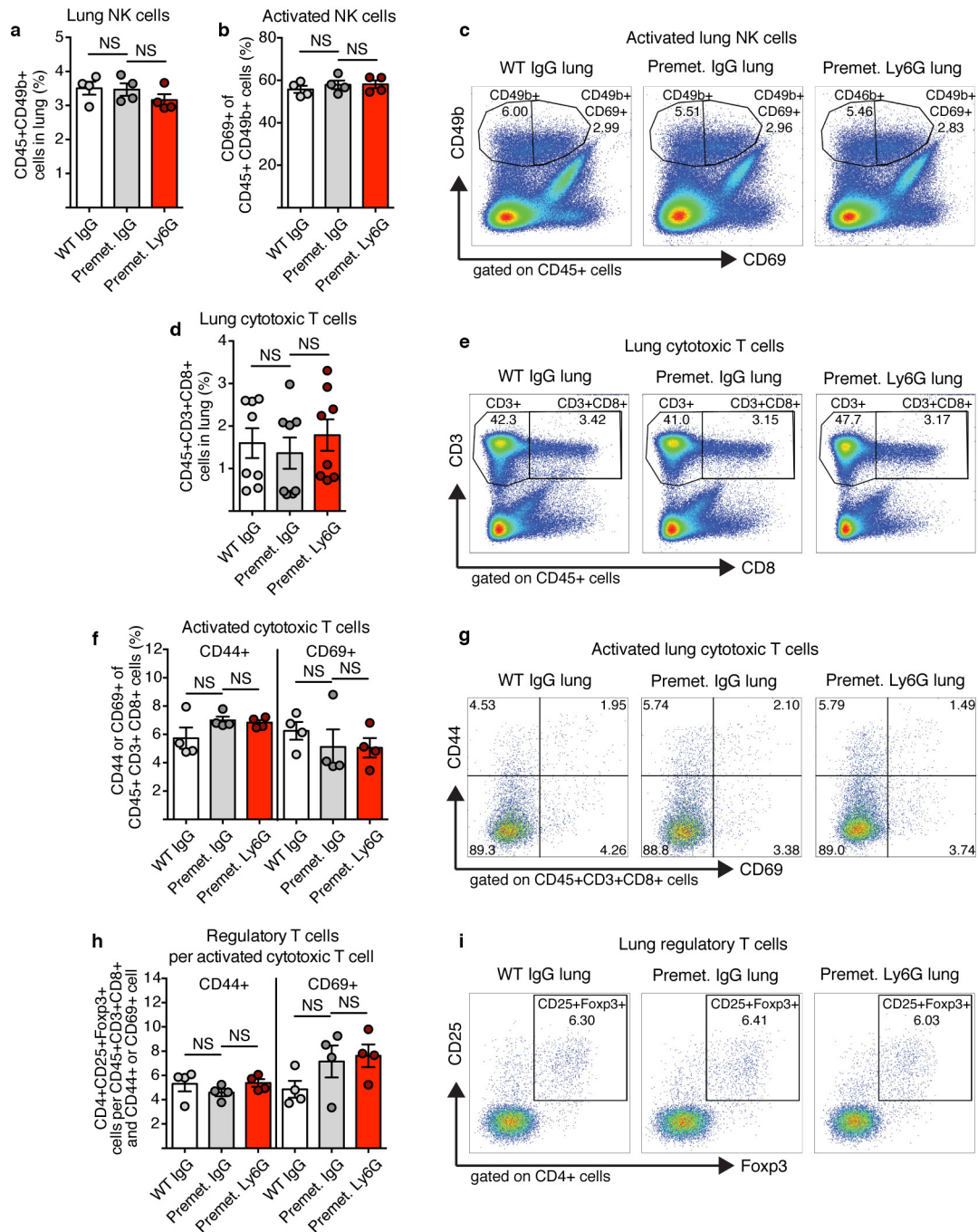
Extended Data Figure 4 | See next page for caption.



**Extended Data Figure 4 | Immune cell frequencies and activation in the pre-metastatic lung of MMTV-PyMT tumour-bearing mice is not dependent on neutrophil presence (part 1).** **a**, Representation of timing and dynamics of neutrophil and cancer cell infiltration into the lung of mice grafted with two mammary tumours by orthotopic injection of  $10^6$  MMTV-PyMT tumour cells. **b–o**, Flow cytometric quantification and representative analysis of the following immune cell types in wild-type (WT) or pre-metastatic (Pre-met.) lungs treated daily with either control IgG or anti-Ly6G (1A8) neutrophil-blocking antibody from tumour onset

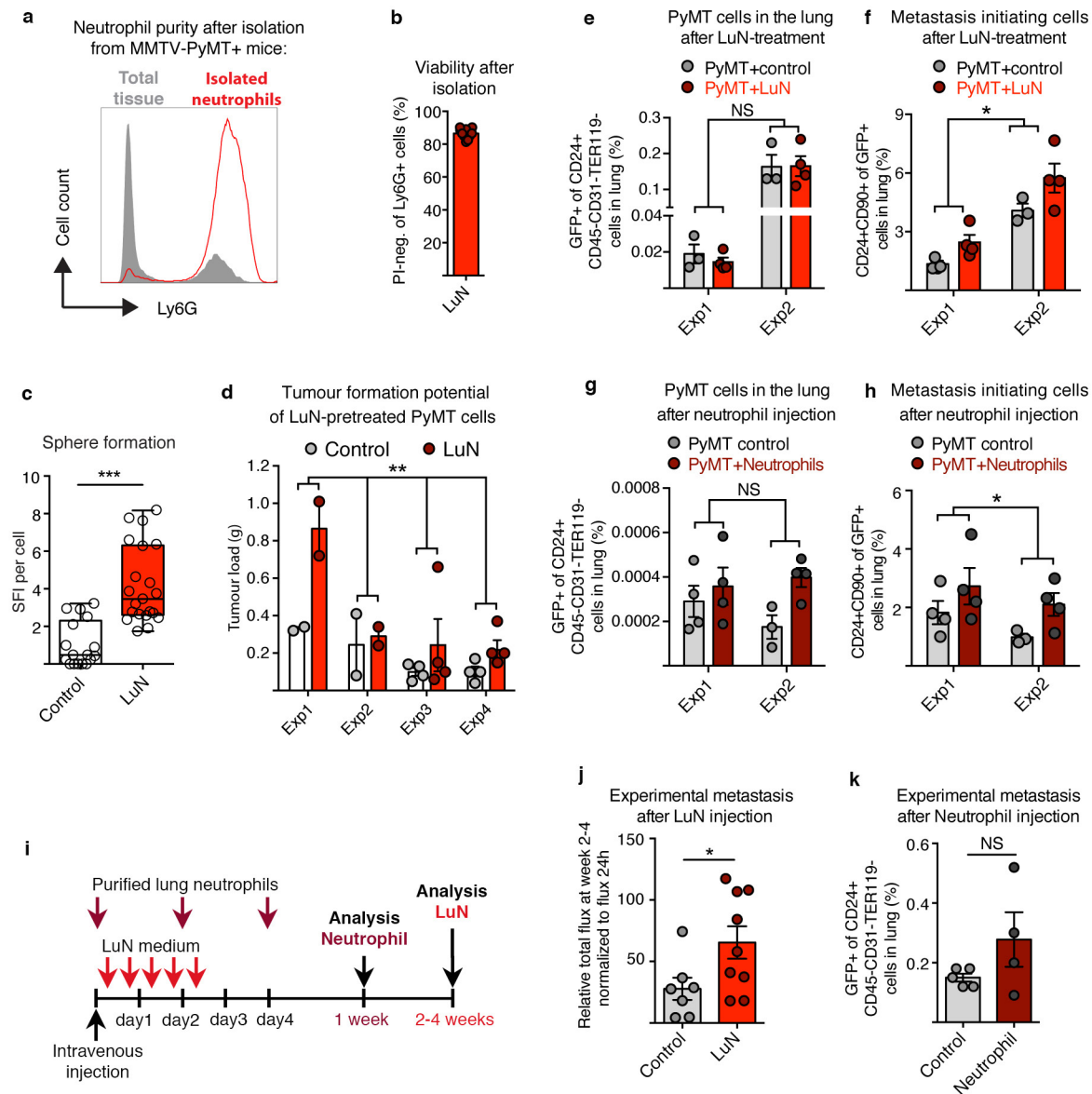
onwards ( $n = 4$  per group if not otherwise indicated): **b, f**, total  $CD45^+$  immune cells ( $n = 12$  per group); **c, g**,  $CD11b^+Ly6G^+$  neutrophils ( $n = 8$  per group); **d, g**,  $CD11b^+SiglecF^+$  eosinophils; **e, g**,  $CD11b^{low}F4/80^{high}$  alveolar macrophages and  $CD11b^{high}F4/80^{low}$  interstitial macrophages; **h, j**,  $CD45^+CD11c^+$  dendritic cells; **i, k**,  $MHC-II^+CD86^+$  activated dendritic cells; **l, n**,  $CD45^+CD19^+$  B cells; and **m, o**,  $MHC-II^+CD86^+$  activated B cells. Statistical analysis by two-sided *t*-test. Data are represented as mean  $\pm$  s.e.m. NS, not significant,  $*P < 0.05$ ,  $**P < 0.01$ ,  $***P < 0.001$ .

Addressing the immunologic response in tumour-bearing mice:



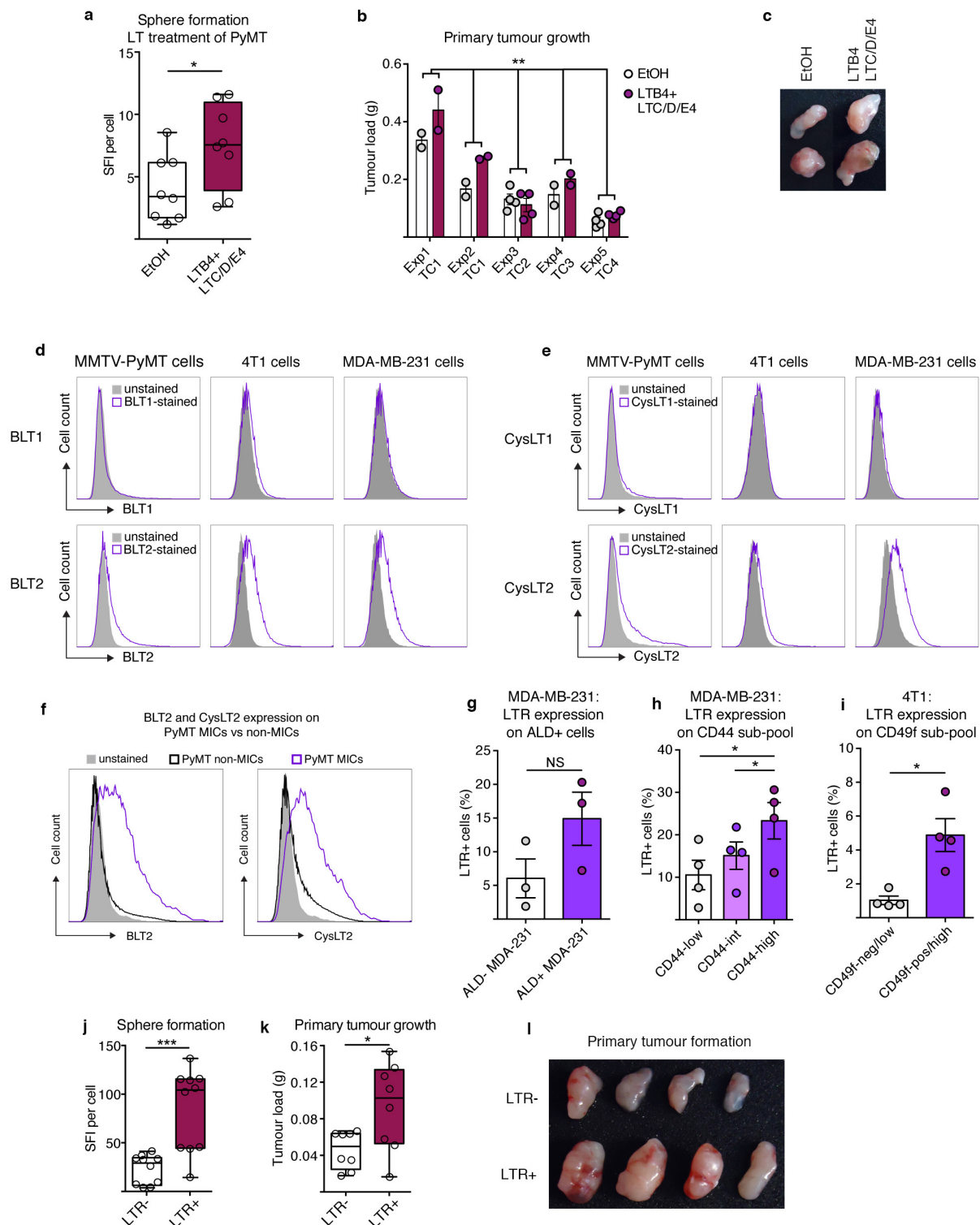
**Extended Data Figure 5 | Immune cell frequencies and activation in the pre-metastatic lung of MMTV-PyMT tumour-bearing mice is not dependent on neutrophil presence (part 2).** a–i, Flow cytometric quantification and representative analysis of the following immune cell types in wild-type (WT) or pre-metastatic (Pre-met.) lungs treated daily with either control IgG or anti-Ly6G (1A8) neutrophil-blocking

antibody from tumour onset onwards ( $n = 4$  per group if not otherwise indicated): a, c, CD45<sup>+</sup>CD49b<sup>+</sup> NK cells; b, c, CD69<sup>+</sup> activated NK cells; d, e, CD45<sup>+</sup>CD3<sup>+</sup>CD8<sup>+</sup> cytotoxic T cells ( $n = 8$  per group); f, g, CD44<sup>+</sup> or CD69<sup>+</sup> activated T cells; and h, i, the ratio of CD4<sup>+</sup>CD25<sup>+</sup>Foxp3<sup>+</sup> regulatory T cells per activated T cell. Statistical analysis by two-sided *t*-test. Data are represented as mean  $\pm$  s.e.m. NS, not significant.



**Extended Data Figure 6 | Neutrophil isolation from the lung of MMTV-PyMT<sup>+</sup> mice and effect of neutrophil-derived factors on tumour formation potential.** **a**, Representative flow cytometric analysis of neutrophil purity after isolation from the pre-metastatic lung compared to total lung tissue. Only neutrophil purity of  $\geq 90\%$  was used for further experiments. **b**, Neutrophil viability was assessed by flow cytometry for propidium iodide (PI) negativity after isolation ( $n = 10$ ). **c**, **d**, MMTV-PyMT cells grown in control or LuN medium for 3 days in adherent conditions were plated in non-attachment conditions followed by sphere quantification at day 10 post-seeding (technical replicate  $n = 17$  (control),  $n = 21$  (LuN) of biological triplicates) (**c**) or  $10^4$  cells grafted onto the mammary gland of *Rag1*-null mice for analysis of tumour formation potential (**d**). Tumour burden was determined by weighing about 3 weeks after ( $n = 12$  per group), complementary to Fig. 2d. **e-h**, Flow cytometric quantification of frequencies of total present GFP-labelled MMTV-PyMT cells (**e**, **g**) and frequencies of CD24<sup>+</sup>CD90<sup>+</sup> MICs among total GFP-labelled MMTV-PyMT cells (**f**, **h**) in the lung of *Rag1*-null mice 3 days

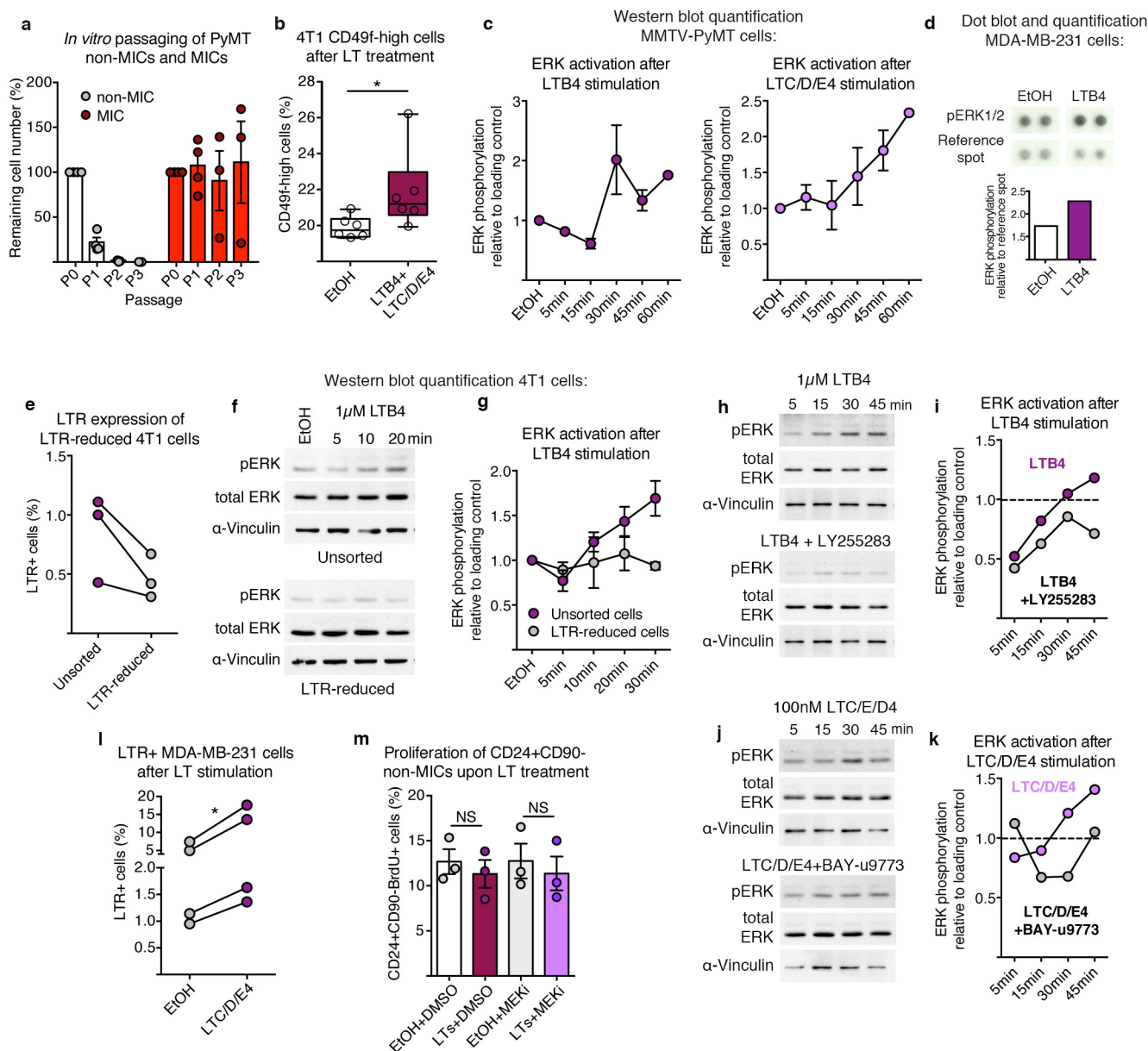
after intravenous injection of  $5 \times 10^5$  total GFP-labelled MMTV-PyMT cells followed by either three intravenous injections with control or LuN medium ( $n = 6$  (PyMT+control),  $n = 8$  (PyMT+LuN)) (**e**, **f**) or by one intravenous injection with  $25 \times 10^6$  neutrophils freshly isolated from a pre-metastatic lung ( $n = 7$  (PyMT+control),  $n = 8$  (PyMT+neutrophils)) (**g**, **h**). **f**, **h**, Two independent experiments are shown to complement Fig. 2h, i. Exp, experiment. **i-k**, Experimental setup (**i**): *Rag1*-null mice were intravenously injected with  $1-10 \times 10^5$  (**j**) or  $0.5 \times 10^6$  total GFP-labelled MMTV-PyMT cells (**k**) followed by either 3-5 intravenous injections with 200  $\mu$ l control or LuN medium (**j**) or by 3 intravenous injections with  $25 \times 10^6$  neutrophils (**k**) freshly isolated from a pre-metastatic lung. Quantification of experimental metastatic incidence by determination of bioluminescence intensity ( $n = 7$  (control),  $n = 9$  (LuN)) (**j**) or flow cytometric analysis of GFP<sup>+</sup> cancer cells in the lung ( $n = 5$  (control),  $n = 4$  (neutrophil)) (**k**) is shown. Statistical analysis by two-sided *t*-test (**c**, **j**, **k**) and two-way ANOVA (**d-h**). Data are represented as mean  $\pm$  s.e.m. NS, not significant, \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .



**Extended Data Figure 7 | LTRs are expressed on mouse and human breast cancer cells and enriched on metastasis-initiating and highly tumorigenic cancer cell sub-pools.** **a**, Sphere formation potential of MMTV-PyMT cells under presence of LTB4 or LTC/D/E4 (technical replicate  $n = 8$  per group of biological triplicates). **b**, **c**, Three-day LTB4 and LTC/D/E4-treated MMTV-PyMT cells in adherent culture were analysed for primary tumour initiation potential by orthotopic transplantation of  $10^4$  cells in *Rag1*-null mice ( $n = 14$  per group) (**b**). Exp, experiment; TC, tumour cell isolation. Representative image of tumours is shown (**c**). **d**, **e**, Flow cytometric analysis of primary MMTV-PyMT cancer cells, the mouse mammary cancer cell line 4T1 and the human breast cancer cell line MDA-MB-231 for expression of BLT1 or BLT2 (**d**) as well as CysLT1 or CysLT2 (**e**). **f**, Representative flow cytometric analysis of BLT2<sup>+</sup> and CysLT2<sup>+</sup> cells among MMTV-PyMT non-MICs and

MICs. **g–i**, Flow cytometric quantification of LTR expression on Aldefluor (ALD)<sup>+</sup> ( $n = 3$  per group) (**g**) or CD44<sup>high</sup> MDA-MB-231 cells ( $n = 4$  per group) (**h**) as well as CD49f<sup>+/high</sup> 4T1 cells ( $n = 4$  per group) (**i**). **j–l**, Sorted LTR<sup>+</sup> or LTR<sup>-</sup> MMTV-PyMT tumour cells were plated in non-attachment conditions followed by sphere quantification at day 10 post-seeding (technical replicate  $n = 10$  per group of biological duplicates) (**j**) or  $10^3$  cells grafted onto the mammary gland of *Rag1*-null mice for analysis of tumour formation potential. Tumour burden was determined by weighing ( $n = 8$  per group) after 3 weeks (**k**) and representative image of tumours is shown (**l**). Statistical analysis by two-sided *t*-test (**a**, **h–k**) and two-way ANOVA (**b**). Data are represented as mean  $\pm$  s.e.m. NS, not significant; \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

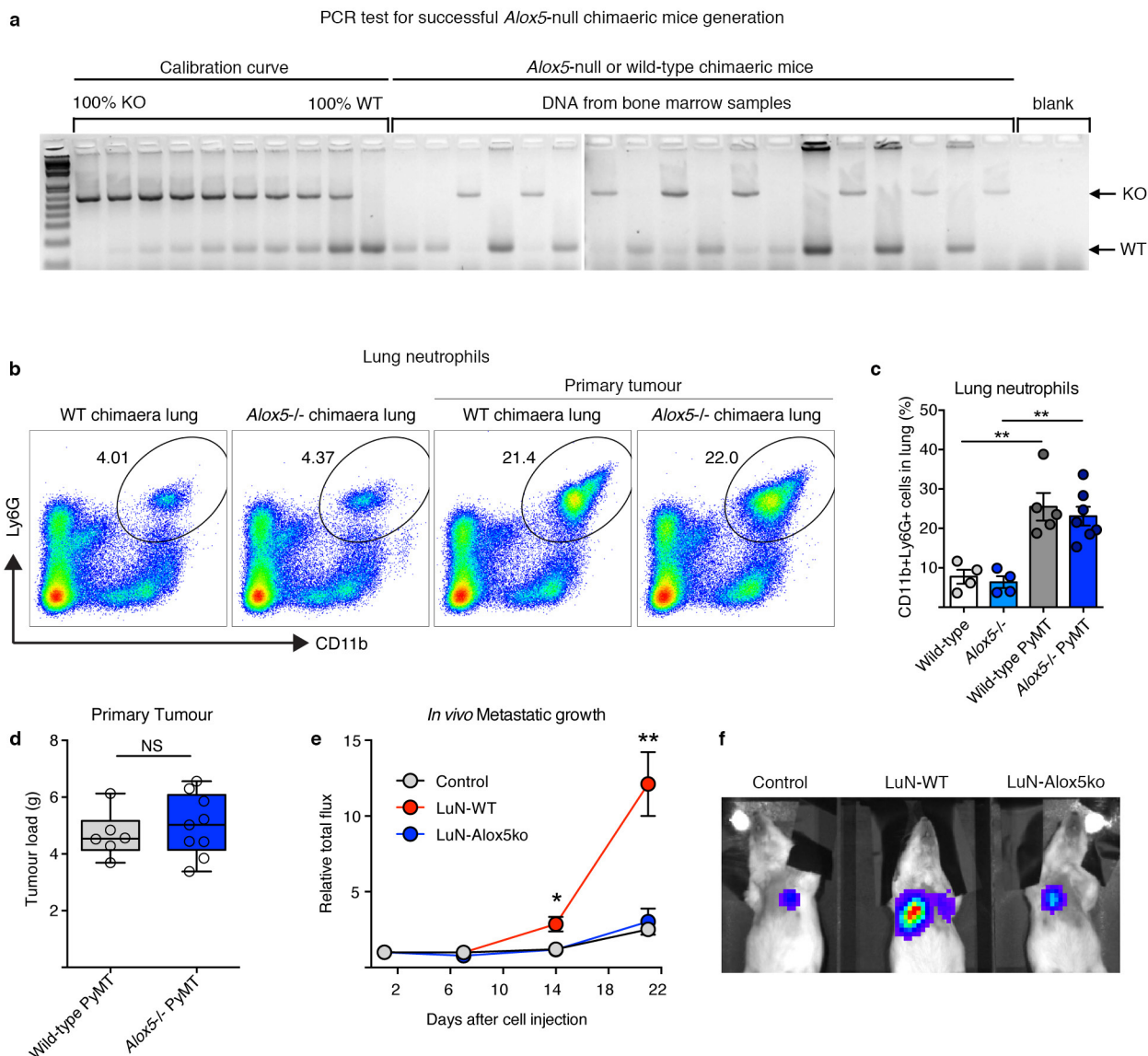




### Extended Data Figure 8 | LTs promote stemness within the total cancer cell population by specifically promoting proliferation of MICs.

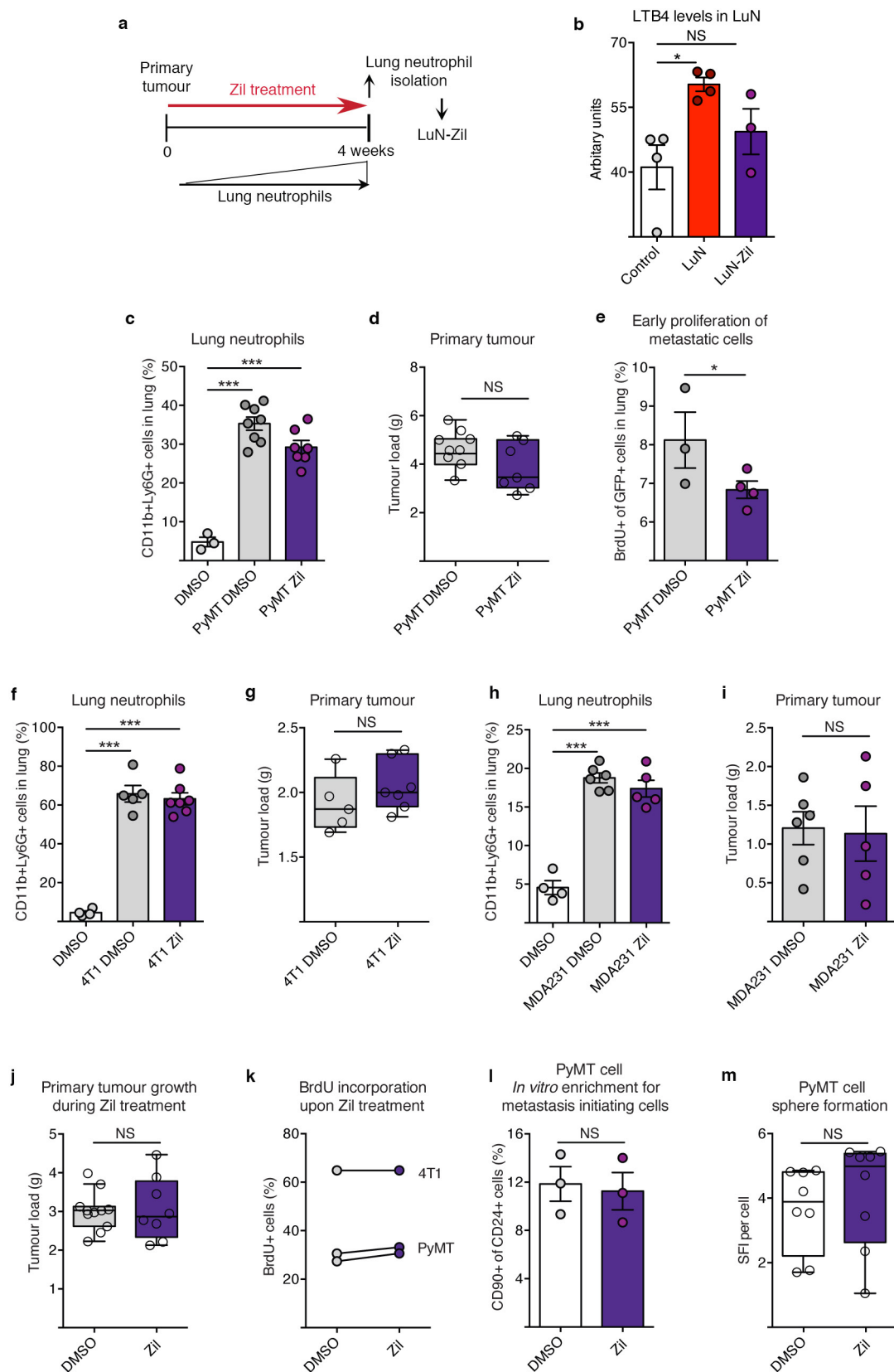
**a**, *In vitro* passing (P indicates passage number) in non-adherent conditions of sorted CD24<sup>+</sup>CD90<sup>+</sup> MICs and CD24<sup>+</sup>CD90<sup>-</sup> non-MICs ( $n = 4$  per group for P0+P1 and  $n = 3$  per group for P2+P3). Quantification was performed by determination of percentage of remaining cell number after 7–10 days. **b**, Flow cytometric quantification of 3-day LT-treated 4T1 cells for frequency of highly tumorigenic CD49f<sup>high</sup> cells ( $n = 6$ ). **c**, Quantification of western blots for ERK1/2 phosphorylation of MMTV-PyMT cells following LTB4 (left) or LTC/D/E4 (right) stimulation relative to  $\alpha$ -vinculin as shown in Fig. 3i ( $n = 2$  per time point except  $n = 9$  (30 min LTB4)). **d**, Dot blot and quantification of ERK1/2 phosphorylation in MDA-MB-231 cells after 3 h stimulation with LTB4 measured by R&D Proteome Profiler Human Phospho-Kinase Array (ARY003B; one-membrane array). **e**, Flow cytometric quantification of LTR expression of sorted LTR-reduced 4T1 cells ( $n = 3$  per group). **f**, **g**, Representative analysis and quantification of western blots for total ERK1/2 and ERK1/2 phosphorylation relative to  $\alpha$ -vinculin of unsorted 4T1 cells or 4T1 cells sorted for LTR negativity ( $n = 2$  per group).

**h–k**, Analysis and quantification of western blot for total ERK1/2 and ERK1/2 phosphorylation relative to  $\alpha$ -vinculin of 4T1 cells following LTB4 (**h**, **i**) or LTC/D/E4 (**j**, **k**) stimulation in the presence of BLT2 inhibitor LY255283 or CysLT2 inhibitor BAY-u9773, respectively (one time series). Dotted lines indicate the control level of ERK1/2 phosphorylation. The decrease of ERK1/2 phosphorylation observed after 5–15 min when adding both leukotrienes and their receptor inhibitors is due to the increase in ethanol concentration. Data are shown as ERK1/2 phosphorylation recovery and increase from 5 to 45 min after stimulation (**i**, **k**). **l**, Flow cytometric quantification of 3-day LTC/D/E4-treated MDA-MB-231 cells for frequency of LTR<sup>+</sup> cells ( $n = 4$  per group). **m**, Three-day LT-treated MMTV-PyMT cells in adherent culture were analysed for BrdU incorporation of CD24<sup>+</sup>CD90<sup>-</sup> non-MICs in the additional presence of PD0325901 MEK inhibitor (MEKi;  $n = 3$  per group). DMSO, dimethylsulfoxide treated; EtOH, ethanol treated. Statistical analysis by two-sided *t*-test (**l**, **m**), and one-sided *t*-test (**b**). Data are represented as mean  $\pm$  s.e.m. NS, not significant; \* $P < 0.05$ . Blot source data are shown in Supplementary Fig. 1.



**Extended Data Figure 9 | Analysis of *Alox5*-null bone marrow chimaeric mice transplanted with primary mammary MMTV-PyMT tumours and failure of *Alox5*-null neutrophils to support cancer cell metastatic initiation potential.** **a**, Efficiency of chimaeric mice generation was determined by semi-quantitative PCR analysis of DNA isolated from the bone marrow of lethally irradiated wild-type mice reconstituted with wild-type or *Alox5*-null bone marrow. A calibration curve of the ratio between the PCR band amplified from the wild-type (WT) and *Alox5*-null (KO) allele was used to calculate the percentage of bone marrow reconstitution efficiency. Tests of 8 representative *Alox5*<sup>-/-</sup> chimaeric mice and 10 controls are shown. Only mice with >80% *Alox5*-null bone marrow reconstitution were used for further experiments. **b–d**, Analysis of wild-type and *Alox5*-null bone marrow chimaeric mice 1.5 months after transplantation with 2 mammary MMTV-PyMT tumours (10<sup>6</sup> PyMT

cells) or tumour-free controls. Representative flow cytometric analysis (**b**) and quantification of CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophil presence in the lung (**c**) ( $n = 4$  (wild type),  $n = 4$  (*Alox5*<sup>-/-</sup>),  $n = 5$  (wild-type PyMT),  $n = 7$  (*Alox5*<sup>-/-</sup> PyMT)) as well as primary mammary tumour burden ( $n = 6$  (wild-type PyMT),  $n = 9$  (*Alox5*<sup>-/-</sup> PyMT)) (**d**). **e**, **f**,  $5 \times 10^5$  luciferase-expressing MMTV-PyMT cells treated with control, wild-type LuN (LuN-WT) or *Alox5*-deficient neutrophil-derived LuN (LuN-Alox5ko) medium for 3 days in adherent culture were intravenously injected into *Rag1*-null mice. Quantification of cancer-cell-derived bioluminescence in the lung over time ( $n = 5$  (control),  $n = 5$  (LuN-WT),  $n = 4$  (LuN-Alox5ko)) (**e**) and representative image is shown (**f**). Statistical analysis by two-sided *t*-test. Data are represented as mean  $\pm$  s.e.m. \* $P < 0.05$ , \*\* $P < 0.01$ . Blot source data are shown in Supplementary Fig. 2.



Extended Data Figure 10 | See next page for caption.

**Extended Data Figure 10 | Breast cancer cell growth, proliferation and self-renewal are not directly affected by treatment with the Alox5 inhibitor Zil.** **a, b**, Neutrophils were isolated from the lungs of MMTV-PyMT mammary tumour-bearing mice treated daily with Zil and used to condition culture medium (LuN-Zil) (**a**). Enzyme-immunoassay analysis of LTB<sub>4</sub> levels in control, LuN or LuN-Zil medium ( $n = 4$  (control),  $n = 4$  (LuN),  $n = 3$  (LuN-Zil)) (**b**). **c, d, f–i**, Analysis of CD11b<sup>+</sup>Ly6G<sup>+</sup> neutrophils in the lung by flow cytometry (**c, f, h**) and primary tumour burden (**d, g, i**) at the time of analysis of *Rag1*-null mice orthotopically transplanted and intravenously injected with GFP-labelled 10<sup>5</sup> primary MMTV-PyMT cancer cells ( $n = 3$  (DMSO),  $n = 9$  (PyMT DMSO),  $n = 7$  (PyMT Zil)) (**c, d**), 10<sup>5</sup> mouse 4T1 cancer cells ( $n = 4$  (DMSO),  $n = 5$  (4T1 DMSO),  $n = 7$  (4T1 Zil)) (**f, g**) or 10<sup>6</sup> human MDA-MB-231 cancer cells ( $n = 4$  (DMSO),  $n = 6$  (MDA231 DMSO),  $n = 5$  (MDA231 Zil)) (**h, i**), and treated with Zil to complement Fig. 4d–k. **e**, Determination of *in vivo* cancer cell proliferation 18 h after intravenous injection of 10<sup>5</sup> GFP-labelled MMTV-PyMT cancer cells into MMTV-PyMT

tumour-bearing, Zil-treated mice by 6 h BrdU pulse and flow cytometric quantification of BrdU<sup>+</sup> among GFP<sup>+</sup> cancer cells in the lung ( $n = 3$  (PyMT DMSO),  $n = 4$  (PyMT Zil)). **j**, Quantification of mammary tumour load of control (DMSO) or Zil-treated wild-type mice 4–6 weeks after orthotopic transplantation with 10<sup>6</sup> MMTV-PyMT cells onto the mammary gland. Daily Zil treatment started 1 day prior to mammary tumour engraftment ( $n = 11$  (DMSO),  $n = 8$  (Zil)). **k**, Flow cytometric quantification of BrdU incorporation after a 3 h pulse of two primary MMTV-PyMT cell preparations and one culture of the mouse 4T1 cell line treated with 1  $\mu$ M Zil for 24 h in adherent conditions. **l**, Flow cytometric quantification of frequency of CD24<sup>+</sup>CD90<sup>+</sup> MICs in total MMTV-PyMT cells after 3-day treatment with 1  $\mu$ M Zil or control DMSO in adherent culture ( $n = 3$  per group). **m**, Sphere formation of MMTV-PyMT cancer cells in the presence of 1  $\mu$ M Zil after 7 days (technical replicate  $n = 8$  per group of biological duplicates). Statistical analysis by two-sided *t*-test (**b–d, f–m**) and one-sided *t*-test (**e**). Data are represented as mean  $\pm$  s.e.m. NS, not significant, \* $P < 0.05$ , \*\*\* $P < 0.001$ .



# Genetic predisposition to neuroblastoma mediated by a *LMO1* super-enhancer polymorphism

Derek A. Oldridge<sup>1,2\*</sup>, Andrew C. Wood<sup>3\*</sup>, Nina Weichert-Leahey<sup>4,5</sup>, Ian Crimmins<sup>1</sup>, Robyn Sussman<sup>1</sup>, Cynthia Winter<sup>1</sup>, Lee D. McDaniel<sup>1</sup>, Maura Diamond<sup>1</sup>, Lori S. Hart<sup>1</sup>, Shizhen Zhu<sup>6</sup>, Adam D. Durbin<sup>4,5</sup>, Brian J. Abraham<sup>7</sup>, Lars Anders<sup>7</sup>, Lifeng Tian<sup>8</sup>, Shile Zhang<sup>9</sup>, Jun S. Wei<sup>9</sup>, Javed Khan<sup>9</sup>, Kelli Bramlett<sup>10</sup>, Nazneen Rahman<sup>11</sup>, Mario Capasso<sup>12,13</sup>, Achille Iolascon<sup>12,13</sup>, Daniela S. Gerhard<sup>14</sup>, Jaime M. Guidry Auvil<sup>14</sup>, Richard A. Young<sup>7</sup>, Hakon Hakonarson<sup>8,15</sup>, Sharon J. Diskin<sup>1,15,16</sup>, A. Thomas Look<sup>4,5</sup> & John M. Maris<sup>1,15,16</sup>

Neuroblastoma is a paediatric malignancy that typically arises in early childhood, and is derived from the developing sympathetic nervous system. Clinical phenotypes range from localized tumours with excellent outcomes to widely metastatic disease in which long-term survival is approximately 40% despite intensive therapy. A previous genome-wide association study identified common polymorphisms at the *LMO1* gene locus that are highly associated with neuroblastoma susceptibility and oncogenic addiction to *LMO1* in the tumour cells<sup>1</sup>. Here we investigate the causal DNA variant at this locus and the mechanism by which it leads to neuroblastoma tumorigenesis. We first imputed all possible genotypes across the *LMO1* locus and then mapped highly associated single nucleotide polymorphism (SNPs) to areas of chromatin accessibility, evolutionary conservation and transcription factor binding sites. We show that SNP rs2168101 G>T is the most highly associated variant (combined  $P = 7.47 \times 10^{-29}$ , odds ratio 0.65, 95% confidence interval 0.60–0.70), and resides in a super-enhancer defined by extensive acetylation of histone H3 lysine 27 within the first intron of *LMO1*. The ancestral G allele that is associated with tumour formation resides in a conserved GATA transcription factor binding motif. We show that the newly evolved protective TATA allele is associated with decreased total *LMO1* expression ( $P = 0.028$ ) in neuroblastoma primary tumours, and ablates GATA3 binding ( $P < 0.0001$ ). We demonstrate allelic imbalance favouring the G-containing strand in tumours heterozygous for this SNP, as demonstrated both by RNA sequencing ( $P < 0.0001$ ) and reporter assays ( $P = 0.002$ ). These findings indicate that a recently evolved polymorphism within a super-enhancer element in the first intron of *LMO1* influences neuroblastoma susceptibility through differential GATA transcription factor binding and direct modulation of *LMO1* expression in *cis*, and this leads to an oncogenic dependency in tumour cells.

Genome-wide association study (GWAS) efforts frequently identify highly statistically significant genetic associations within non-coding regulatory regions of the genome, but the underlying causal DNA sequence variations have only been identified in a few instances. A neuroblastoma GWAS has identified several disease susceptibility loci<sup>1–7</sup>, with the signal within the LIM domain only 1 (*LMO1*) locus at 11p15 (ref. 1), a transcriptional co-regulator containing two zinc finger LIM domains that nucleate and regulate transcription factor

complexes, being most robust. The main members of the LMO gene family, *LMO1–4*, are all implicated in cancer including *LMO1* and *LMO2* translocations in T-cell leukaemia<sup>8</sup>, and we previously provided the first evidence that *LMO1* was a bona fide neuroblastoma oncogene<sup>1</sup>. Here, we sought to identify the causal polymorphism(s) driving the *LMO1* genetic association with neuroblastoma susceptibility as a basis for understanding neuroblastoma initiation and addiction mechanisms.

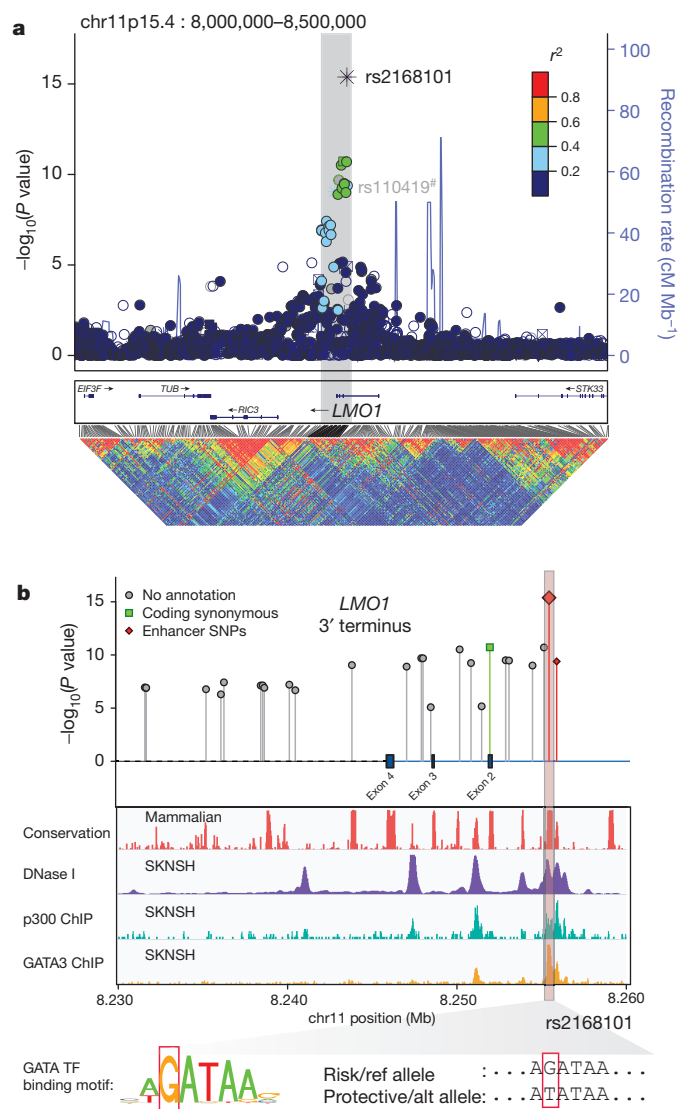
We first performed fine mapping of associated germline SNPs and indels at the *LMO1* gene locus by imputation to the 1000 Genomes Project for our European-American neuroblastoma GWAS<sup>6</sup>. This identified 27 SNPs with minor allele frequency (MAF) >0.01 and an association  $P < 1 \times 10^{-5}$  (Fig. 1a and Extended Data Table 1). We further prioritized associated variants by evolutionary conservation, and by their regulatory potential inferred through neuroblastoma-specific DNase I hypersensitivity mapping and chromatin immunoprecipitation sequencing (ChIP-seq) from the ENCODE Consortium (Fig. 1b). These data showed that the most significantly associated SNP at the *LMO1* locus (rs2168101, odds ratio = 0.67,  $P = 4.14 \times 10^{-16}$ ) resides within a highly conserved and active enhancer region inferred by DNase I sensitivity and p300 binding in the SKNSH neuroblastoma cell line (Fig. 1b). Notably, we found no rare or common non-synonymous coding variants in *LMO1* in a combined cohort of 482 unique neuroblastoma cases with germline whole-genome ( $n = 136$ ), whole-exome ( $n = 222$ ) and/or targeted DNA sequencing ( $n = 183$ ) (see Extended Data Table 2 and Supplementary Data).

Because rs2168101 genotypes were imputed in our analyses (Extended Data Fig. 1), we next directly genotyped this SNP in 146 out of 2,101 European-American cases, and measured an 86% imputation accuracy (Supplementary Table 1). We additionally directly genotyped rs2168101 in two independent cohorts from the UK and Italy, with both showing robust replication (Table 1). We did not observe replication in an independent African-American cohort. Notably, the protective T allele is common in Europeans (CEU HapMap: 28%) and East Asians (CHB+JPT HapMap: 32%), but is rare or absent in Africans, indicating recent expansion of the rs2168101 protective allele in non-African human populations. Meta-analysis demonstrated a combined association  $P = 7.47 \times 10^{-29}$  across 8,553 controls and 3,254 cases (Table 1).

As causal SNPs driving GWAS associations may disrupt transcription factor binding at distal enhancers, we sought to identify candidate SNPs disrupting known JASPAR motifs<sup>9</sup>, which revealed that lead

<sup>1</sup>Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. <sup>2</sup>Medical Scientist Training Program, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>3</sup>Department of Molecular Medicine and Pathology, University of Auckland, Auckland, Auckland Region 1142, New Zealand. <sup>4</sup>Department of Pediatric Oncology, Dana Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02215, USA. <sup>5</sup>Division of Pediatric Hematology/Oncology, Boston Children's Hospital, Boston, Massachusetts 02115, USA. <sup>6</sup>Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, Minnesota 55905, USA. <sup>7</sup>Whitehead Institute for Biomedical Research and MIT, Boston, Massachusetts 02142, USA. <sup>8</sup>Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. <sup>9</sup>Pediatric Oncology Branch, National Cancer Institute, Bethesda, Maryland 20892, USA. <sup>10</sup>Thermo Fisher Scientific, Austin, Texas 78744, USA. <sup>11</sup>The Institute of Cancer Research, London SM2 5NG, UK. <sup>12</sup>University of Naples Federico II, 80131 Naples, Italy. <sup>13</sup>CEINGE Biotechnologie Avanzate, 80131 Naples, Italy. <sup>14</sup>Office of Cancer Genomics, National Cancer Institute, Bethesda, Maryland 20892, USA. <sup>15</sup>Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>16</sup>Abramson Family Cancer Research Institute, Philadelphia, Pennsylvania 19104, USA.

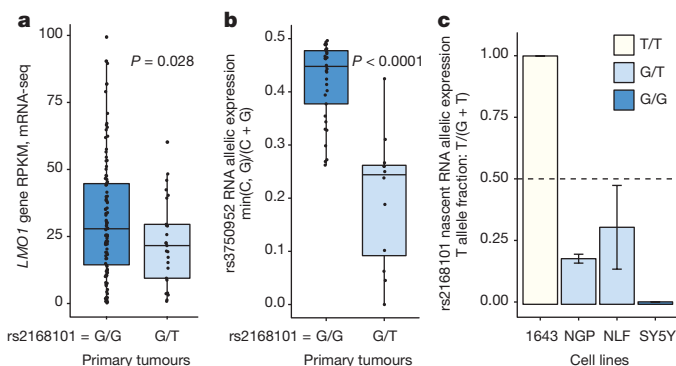
\*These authors contributed equally to this work.



**Figure 1 | Imputation-based GWAS and epigenomic profiling by ENCODE identifies rs2168101 as a candidate functional SNP at *LMO1*.** **a**, Manhattan plot for neuroblastoma GWAS (cases = 2,101; controls = 4,202). The neuroblastoma-associated region falls within a 40-kilobase (kb) haplotype block (grey box) in Europeans, encompassing the *LMO1* 3'-terminus. rs2168101 is the most associated variant and is moderately correlated (maximum  $r^2 = 0.52$ ) with other variants. The sentinel SNP reported previously, rs110419, is also highlighted (#). **b**, Associated variants ( $P < 1 \times 10^{-5}$ ) are plotted with ENCODE tracks for neuroblastoma cell line SKNSH. Two SNPs, rs2168101 and rs7948497, were annotated 'enhancer SNPs' based on overlapping DNase peaks binding p300. The rs2168101 G>T SNP disrupts an evolutionarily conserved GATA transcription factor (TF) motif (5'-A[G/T]ATAA-3'). SKNSH has an rs2168101 = G/G genotype that preserves GATA binding, supported by ENCODE GATA3 ChIP-seq.

candidate SNP rs2168101 resides in a highly conserved GATA-binding motif (5'-A[G/T]ATAA-3', mammalian phastCons score = 100%) (Fig. 1b). ENCODE transcription factor ChIP-seq confirmed GATA2 and GATA3 binding at the rs2168101 GATA motif in the neuroblastoma cell lines SKNSH and SHSY5Y, which are G/G homozygous, thereby preserving the GATA motif (Fig. 1b). No other associated variant showed this unique combination of evolutionary conservation, active enhancer localization, and disruption of a transcription factor binding motif, including the sentinel SNP rs110419 ( $P = 1.17 \times 10^{-13}$ ) from our original report<sup>1</sup>.

To test for the possibility of multiple statistical signals or enhancers not marked by conservation or p300 at the *LMO1* locus, we repeated association testing conditional on imputed rs2168101 genotypes and

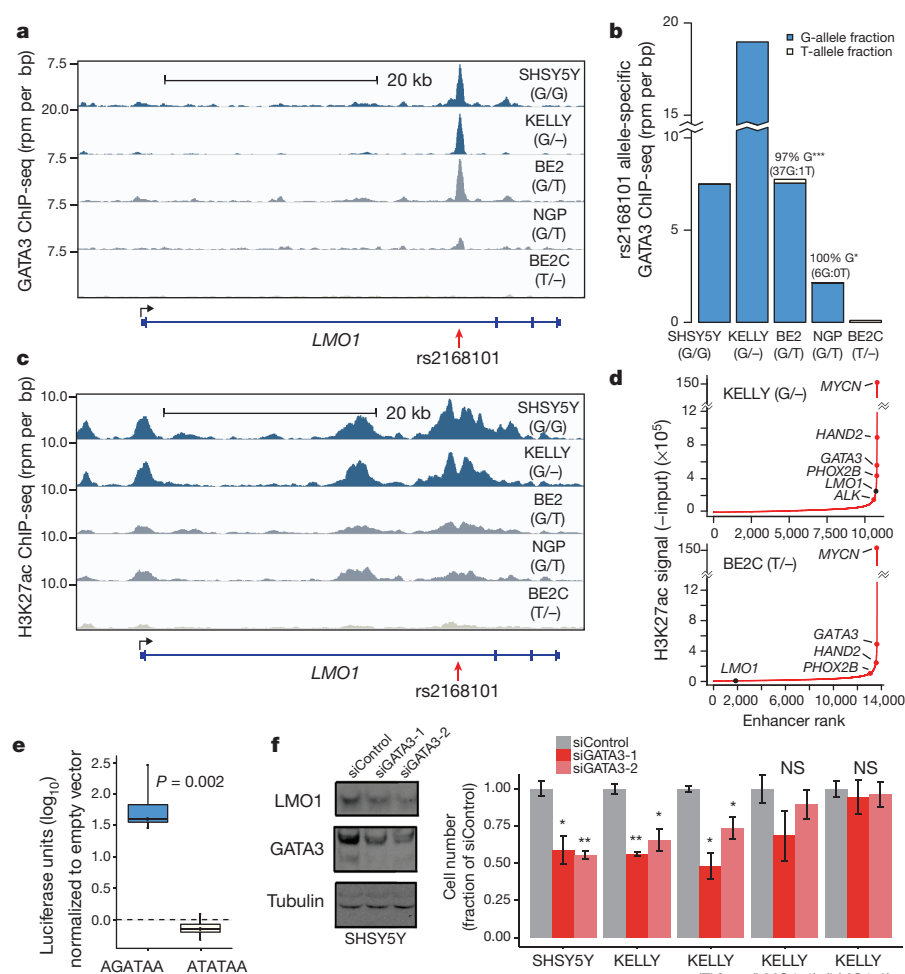


**Figure 2 | RNA expression of *LMO1* associates with rs2168101 genotype consistent with regulation in *cis*.** **a**, mRNA-seq across 127 primary tumours genotyped for rs2168101 (G/G = 102, G/T = 25, T = 0) revealed a significant decrease in *LMO1* gene expression between G/T and G/G tumours ( $t$ -test  $P = 0.028$ ). RPKM, reads per kilobase per million reads. **b**, Using the synonymous exonic SNP, rs3750952, to measure allelic expression by mRNA-seq revealed significantly more allelic imbalance in 12 heterozygous neuroblastoma tumours (rs2168101 = G/T) than in 33 homozygous tumours (rs2168101 = G/G) ( $t$ -test  $P = 5.3 \times 10^{-5}$ ). **c**, Allelic expression for rs2168101 from targeted nascent RNA-seq in four neuroblastoma cell lines. The two heterozygous cell lines (rs2168101 = G/T) exhibited significantly reduced T-allele expression compared to the G allele ( $t$ -test  $P = 1.6 \times 10^{-4}$  and  $1.5 \times 10^{-2}$  for NGP and NLF, respectively; error bars denote 95% confidence intervals across  $n = 3$  duplicate experiments).

observed no significant variants after multiple test correction (most significant variant: rs34544683, nominal  $P = 9.0 \times 10^{-4}$ , Bonferroni  $P = 1$ ; Extended Data Fig. 2a). To test whether the rs2168101 signal can be equally captured by other variants, we also performed reciprocal association tests for rs2168101 conditioned on all 27 other SNPs within 1.5 megabases (Mb) of *LMO1* passing thresholds MAF > 0.01 and nominal  $P < 1 \times 10^{-5}$ . Notably, rs2168101 remained significant across all conditional tests (worst-case nominal  $P = 2.6 \times 10^{-7}$ , Bonferroni  $P = 0.002$ ; Extended Data Fig. 2b). These results are consistent with a single underlying signal at the *LMO1* locus, and re-affirm that rs2168101 is the single best causal SNP candidate, because its association with neuroblastoma cannot be accounted for by other variants.

We next sought to determine whether rs2168101 genotypes were associated with *LMO1* expression by messenger RNA sequencing (mRNA-seq) of 127 primary high-risk neuroblastoma tumours. Genotyping rs2168101 yielded 102 G/G, 25 G/T and no T/T tumours (MAF = 9.8%). We observed significantly higher *LMO1* expression in G/G versus G/T genotype tumours ( $t$ -test  $P = 0.028$ ; Fig. 2a). Notably, the absence of protective homozygous T/T genotypes in this high-risk neuroblastoma cohort is consistent with our previous observation that the risk alleles predispose to the high-risk phenotypic subset<sup>1</sup> (for clinical covariate associations, see Extended Data Table 3). Accordingly, the rs2168101 G/G genotype is highly associated with decreased neuroblastoma patient event-free ( $P = 0.0004$ ) and overall ( $P = 0.0004$ ) survival compared to G/T and T/T genotypes together in our European-American cohort (Extended Data Fig. 3). Two cell lines with homozygous T/T or T/- genotypes expressed *LMO1* at comparatively lower levels than cell lines containing the G allele (Extended Data Fig. 4a).

GATA transcription factors mediate chromatin looping and facilitate long-range enhancer-promoter interactions to regulate target gene expression<sup>10</sup>. We therefore sought to confirm allelic imbalance of *LMO1* transcripts (a hallmark of gene regulation in *cis*), which could result from differential GATA-binding caused by rs2168101. First, because the rs2168101 intronic SNP is not detectable by mRNA-seq, we identified the *LMO1* exonic synonymous SNP, rs3750952, which can measure allelic expression in the heterozygous state. We identified 45 tumours with



**Figure 3 | The rs2168101 protective T allele disrupts GATA3 binding and negatively associates with the LMO1 super-enhancer in neuroblastoma cells.** **a**, Normalized GATA3 ChIP-seq signal at rs2168101 in five neuroblastoma cell lines (rs2168101 genotypes: SHSY5Y = G/G, KELLY = G/-, BE2 = G/T, NGP = G/T, BE2C = T/-). rpm, reads per million. **b**, Allele-specific binding of GATA3 at rs2168101. GATA3 binding highly favoured the risk G allele in heterozygous lines (BE2: 0.97 G-allele fraction from 38 reads, 95% confidence interval: 0.86–1.00, binomial  $P = 2.8 \times 10^{-10}$ ; NGP: 1.00 G-allele fraction from 6 reads, 95% confidence interval: 0.54–1.00, binomial  $P = 0.03$ ). **c**, Normalized ChIP-seq signal for H3K27ac at rs2168101. **d**, Ranked H3K27ac signal across all enhancers in MYCN-amplified KELLY and BE2C lines. Super-enhancers associate with key neuroblastoma genes, highlighted on the curve. There is an LMO1-associated super-enhancer in G-allele-containing lines SHSY5Y, KELLY and BE2, but not in BE2C, which lacks the G allele. **e**, Luciferase reporter assay for LMO1 enhancer region. The risk G allele preserved enhancer activity ( $t$ -test  $P = 0.002$  across  $n = 4$  independent clones, each with  $n = 5$  technical replicates), whereas the protective T allele was indistinguishable from empty vector. **f**, Left, protein blots for GATA3, LMO1 and tubulin in SHSY5Y cells treated with control (siControl) short interfering RNAs (siRNAs), or with siRNAs targeting GATA3 (siGATA3-1 and siGATA3-2), at 72 h post-treatment. Right, cell counts for cell lines SHSY5Y, KELLY, KELLY stably overexpressing control vector (EV) and KELLY with forced LMO1 overexpression (LMO1-1 and LMO1-2) treated with siRNAs at 72 h post-transfection (see Extended Data Fig. 7 for complete growth curves). Rescue of suppressed cell growth after GATA3 depletion by forced LMO1 expression was observed at 72 h. Error bars denote  $\pm$  s.e.m.  $*P < 0.05$ ,  $**P < 0.001$  by  $t$ -test.  $n = 3$  independent transfections,  $n = 9$  technical replicates.

the necessary rs3750952 = C/G genotype, and then directly genotyped rs2168101 (G/G = 33, G/T = 12, T/T = 0) in this panel. By mRNA-seq, there was greater allelic imbalance in 12 tumours that were heterozygous for rs2168101 (G/T) than in 33 homozygous tumours (rs2168101 = G/G;  $t$ -test  $P < 0.0001$ ; Fig. 2b). We next used targeted sequencing of nuclear-enriched nascent RNAs in four neuroblastoma cell lines (G/G = 1, G/T = 2, T/T = 1) to provide direct ascertainment of allele-specific expression at rs2168101. In both heterozygous lines, we observed allelic imbalance that significantly favoured the risk G allele over the protective T allele (Fig. 2c). Collectively, these results indicate that the intact GATA motif at rs2168101 results in significantly higher LMO1 expression levels than the TATA coded by the alternative allele. Allelic imbalance of LMO1 was not driven by somatic DNA alterations (for example, loss of heterozygosity) that could affect allelic dosage (Extended Data Fig. 4b).

Examination of neuroblastoma transcriptome data for 127 primary tumours showed that GATA2 and GATA3 are overexpressed compared to other members of the GATA transcription factor family (Extended Data Fig. 5a), and that GATA3 is the most highly expressed. Additionally, protein immunoblotting showed that GATA3 is uniformly highly expressed in neuroblastoma cell lines, while LMO1 is highly expressed in the G/G (SKNSH and SHSY5Y), G/- (KELLY) and G/T (IMR32) cell lines, but only barely detectable in the BE2C cell line that lacks a G allele at the rs2168101 locus (Extended Data Fig. 5b). We therefore performed ChIP-seq using a GATA3 antibody in neuroblastoma cell lines, and observed robust GATA3 binding at rs2168101 in lines containing the G allele (SHSY5Y, KELLY, BE2 and NGP) but not in a line containing only a T allele (BE2C; Fig. 3a). We then specifically considered GATA3 ChIP-seq reads overlapping rs2168101, and we observed strong preferential binding to the G allele in the G/T

**Table 1 | Replication and meta-analysis of rs2168101 association**

SNP	Ref/alt (major/minor) allele	Cohort	MAF cases	MAF controls	Additive $P$ value	Additive odds ratio	Het odds ratio (GT vs GG)	Hom odds ratio (TT vs GG)
rs2168101	G/T	European-American*	0.242 ( $n = 2,101$ )	0.313 ( $n = 4,202$ )	$4.14 \times 10^{-16}$	0.67 (0.61–0.74)	0.69 (0.62–0.77)	0.52 (0.42–0.64)
		Italian	0.164 ( $n = 420$ )	0.250 ( $n = 751$ )	$2.07 \times 10^{-6}$	0.61 (0.50–0.75)	0.57 (0.44–0.74)	0.40 (0.21–0.75)
		UK	0.190 ( $n = 369$ )	0.311 ( $n = 1,109$ )	$5.86 \times 10^{-10}$	0.56 (0.47–0.68)	0.51 (0.39–0.66)	0.31 (0.18–0.53)
		African-American*	0.0865 ( $n = 364$ )	0.0891 ( $n = 2,491$ )	0.20	0.79 (0.56–1.13)	0.96 (0.71–1.30)	1.07 (0.38–3.04)
		Combined			$7.47 \times 10^{-29}$	0.65 (0.60–0.70)	0.67 (0.61–0.73)	0.49 (0.41–0.59)

Alt, alternative; het, heterozygous; hom, homozygous; ref, reference. Odds ratio 95% confidence intervals are shown in parentheses.

\*Imputed genotypes and correction for population stratification.



heterozygous cell lines BE2 (0.97 G-allele fraction from 38 reads, 95% confidence interval: 0.86–1.00, binomial test  $P = 2.8 \times 10^{-10}$ ) and NGP (1.00 G-allele fraction from 6 reads, 95% confidence interval: 0.54–1.00, binomial test  $P = 0.03$ ; Fig. 3b).

Acetylation of histone H3 at lysine 27 (H3K27ac) is a hallmark of active enhancers<sup>11</sup>, and ChIP-seq analysis of SHSY5Y (G/G; not MYCN amplified), KELLY (G/–; MYCN amplified), BE2 (G/T; MYCN amplified) and NGP (G/T; MYCN amplified) neuroblastoma cells showed extensive H3K27 acetylation in the first intron of *LMO1* across rs2168101, which was not observed in BE2C (T/–; MYCN amplified; Fig. 3c). This region is classified as a super-enhancer in G-allele-containing lines SHSY5Y, KELLY and BE2 based on enhancer clustering and especially high H3K27ac signal (NGP was just below the threshold; see Methods), a pattern also observed for other known oncogenes and tumour suppressor genes in this disease<sup>12</sup> (Fig. 3d and Extended Data Fig. 6a). No super-enhancer was observed in BE2C or Jurkat T-ALL cells that also express *LMO1* (ref. 13), or in other non-neuroblastoma tissues from ENCODE (Fig. 3d and Extended Data Fig. 6b, c). These results are consistent with recent evidence that disease-associated SNPs frequently affect enhancers that are specific to disease-relevant cell lines and tumour histology, and control developmental stage and tissue-specific gene expression<sup>12,14–18</sup>.

We next performed luciferase reporter assays to measure the effect of rs2168101 alleles on enhancer activity. HEK293T cells transfected with constructs containing the risk G allele demonstrated 30–300-fold higher normalized luminescence compared to the T allele ( $t$ -test  $P = 0.002$ , Fig. 3e), whereas luciferase activity of the T allele was not significantly different from empty vector, indicating that the intact GATA motif is required for robust enhancer activity. Finally, knockdown of GATA3 in SHSY5Y and KELLY cells resulted in both decreased *LMO1* protein levels and suppression of cell growth that was rescued by *LMO1* overexpression (Fig. 3f and Extended Data Fig. 7), indicating the central role of GATA3 in regulating *LMO1* expression levels in neuroblastoma.

Taken together, these data demonstrate the underlying molecular mechanism for a highly robust genetic association to neuroblastoma, mediated by a single common causal SNP rs2168101 that disrupts a GATA transcription factor binding site within a tissue-specific super-enhancer element. The rarity or absence of the protective allele in African populations and its relative depletion in African-Americans may partially explain the more aggressive clinical course in African-American children<sup>19</sup>. Moreover, this work further confirms the utility of association studies to define clinically relevant oncogenic pathways. Finally, the dependence of neuroblastoma cells on super-enhancer-mediated *LMO1* expression provides another potential mechanism for the sensitivity of these tumours to inhibitors of the transcriptional machinery such as CDK7 and BET bromodomain proteins<sup>14,16</sup>, demonstrating the potential of translating basic mechanistic insights of tumour initiation towards novel therapeutic strategies.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 7 February; accepted 2 September 2015.**

**Published online 11 November 2015.**

- Wang, K. *et al.* Integrative genomics identifies *LMO1* as a neuroblastoma oncogene. *Nature* **469**, 216–220 (2011).
- Maris, J. M. *et al.* Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.* **358**, 2585–2593 (2008).
- Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987–991 (2009).
- Capasso, M. *et al.* Common variations in *BARD1* influence susceptibility to high-risk neuroblastoma. *Nature Genet.* **41**, 718–723 (2009).
- Nguyễn Lê, B. *et al.* Phenotype restricted genome-wide association study using a gene-centric approach identifies three low-risk neuroblastoma susceptibility loci. *PLoS Genet.* **7**, e1002026 (2011).

- Diskin, S. J. *et al.* Common variation at 6q16 within *HACE1* and *LIN28B* influences susceptibility to neuroblastoma. *Nature Genet.* **44**, 1126–1130 (2012).
- Diskin, S. J. *et al.* Rare variants in *TP53* and susceptibility to neuroblastoma. *J. Natl. Cancer Inst.* **106**, dju047 (2014).
- Matthews, J. M., Lester, K., Joseph, S. & Curtis, D. J. LIM-domain-only proteins in cancer. *Nature Rev. Cancer* **13**, 111–122 (2013).
- Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
- Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233–1244 (2012).
- Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
- Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
- Sanda, T. *et al.* Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer Cell* **22**, 209–221 (2012).
- Chipmuro, E. *et al.* CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. *Cell* **159**, 1126–1139 (2014).
- Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl Acad. Sci. USA* **110**, 17921–17926 (2013).
- Puissant, A. *et al.* Targeting MYCN in neuroblastoma by BET bromodomain inhibition. *Cancer Discov.* **3**, 308–323 (2013).
- Sur, I., Tuupainen, S., Whittington, T., Aaltonen, L. A. & Taipale, J. Lessons from functional analysis of genome-wide association studies. *Cancer Res.* **73**, 4180–4184 (2013).
- Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
- Henderson, T. O. *et al.* Racial and ethnic disparities in risk and survival in children with neuroblastoma: a Children's Oncology Group study. *J. Clin. Oncol.* **29**, 76–82 (2011).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported in part by NIH grants R01-CA124709 (J.M.M.), R01-CA180692 (J.M.M. and A.T.L.), R00-CA151869 (S.J.D.), RC1MD004418 to the TARGET consortium, 1K99CA178189 (S.Z.), T32-HG000046 (D.A.O.), R01-CA109901 (R.A.Y.), the Giulio D'Angio Endowed Chair (J.M.M.), the PressOn Foundation (J.M.M.), Andrew's Army Foundation (J.M.M.), the Abramson Family Cancer Research Institute (J.M.M.), the Brooke Mulford Foundation (J.M.M.), the University of Pennsylvania Genome Frontiers Institute, an Alex's Lemonade Stand Foundation Innovation Award (A.T.L.), young investigator awards from Alex's Lemonade Stand Foundation (S.Z., A.C.W.) and the CureSearch for Children's Cancer Foundation (S.Z.), grant from the German Cancer Aid 110801 (N.W.-L.), St Baldrick's Foundation Fellow award (A.C.W.), George L. Ohrstrom Jr foundation (A.C.W.), Wellcome Trust Senior Investigator Award Ref: 100210/Z/12/Z (N.R.) and NHS funding to the NIHR Biomedical Research Centre at The Royal Marsden and the ICR (N.R.), Fondazione Italiana per la Lotta al Neuroblastoma (M.C.), Associazione Oncologia Pediatrica e Neuroblastoma (M.C.), and Associazione Italiana per la Ricerca sul Cancro (M.C.). We gratefully acknowledge the Children's Oncology Group (COG) for providing the specimens and clinical data from neuroblastoma patients and thank patients and families for participating in the COG, the UK-based Factors Associated with Childhood Tumors (FACT), and Italian cooperative group studies. We thank A. Renwick who performed the Tagman analyses and A. Zachariou for recruiting participants to the FACT study. We thank G. Blobel for scientific advice and discussion, and generously providing equipment and reagents for ChIP experiments, N. Saeki and H. Sasaki for providing the *LMO1* cDNA clone, and Y. Nakatan for providing the lentiviral vector pOZ-FHN.

**Author Contributions** J.M.M. and A.T.L. conceived the study, guided interpretation of results and guided preparation of the manuscript. D.A.O. and A.C.W. performed and/or oversaw most of the experiments, computational analyses and data interpretation. I.C., R.S., C.W., L.S.H., S.Z., N.W.-L., A.D.D., B.J.A., L.A., L.T., K.B. and R.A.Y. performed the genomic and epigenetic experiments and data analysis including DNA sequencing and ChIP sequencing. L.D.M., S.J.D. and H.H. performed the fine mapping and association testing. J.S.W. and J.K. performed the tumour RNA sequencing. N.R. and M.C. performed the validation genotyping and association testing. M.C. and A.I. replicated the SNP association in the Italian cohort. D.A.O. and A.C.W. drafted the manuscript, while A.T.L. and J.M.M. and other authors edited the manuscript.

**Author Information** ChIP-seq data sets are available under Gene Expression Omnibus (GEO) super series GSE65664, and relevant accession numbers are shown in Supplementary Table 2. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M.M. (maris@chop.edu).



## METHODS

No statistical methods were used to predetermine sample size.

**Genotype imputation and association testing.** A primary European-American cohort of 2,101 cases and 4,202 matched controls were assayed with Illumina HumanHap550 v1, Illumina HumanHap550 v3, and Illumina Human610 SNP arrays as previously described<sup>6</sup>. Genotypes were phased using SHAPEIT v2.2.790 and data from 1000 Genomes phase 1 version 3. Subsequently, imputation was performed using IMPUTE2 v2.3.1 for all SNPs and indel variants annotated in the 1000 Genomes phase 1 version 3. Testing for association with neuroblastoma under an additive genetic effect model was performed using the frequentist likelihood score method implemented in SNPTEST v2.4.1. Genotypes for a previously described African-American cohort of 365 cases and 2,491 controls<sup>20</sup> were imputed and tested for neuroblastoma association using the same analytic pipeline. Statistical adjustment for gender was performed in both cohorts. For population stratification adjustment, the first 20 multidimensional scaling (MDS) components were included as covariates in the European-American cohort, while a measure of African admixture as estimated by the ADMIXTURE software program was used in the African-American cohort. Manhattan plots of SNP position and statistical significance were generated using LocusZoom software. Linkage plots were generated by Haploview software based on HapMap CEU individuals (version 3, release 2) using default settings. All research subjects or their guardians provided informed consent for research, and all institutions involved in this research had regulatory approval for human subjects research.

**Prioritization of candidate causal variants.** All SNPs and indels reported in the 1000 Genomes phase 1 version 3 data were considered as candidate causal variants and were ranked based on a combination of (1) neuroblastoma association in the primary European-American cohort, (2) evolutionary conservation, (3) DNase I hypersensitivity, and (4) transcription factor binding motif matching. Neuroblastoma association in European-Americans was evaluated as described above. Conservation scores were computed as the average of the phastCons46way-Placental UCSC conservation track score for all bases from the -10 position to the +10 position surrounding each candidate variant. A DNase I hypersensitivity score was calculated by counting the number of sequencing tags from the -100 position to the +100 position around each candidate variant in ENCODE data for the neuroblastoma cell line, SK-N-SH. Position weight matrices representing transcription factor binding motifs were obtained from the JASPAR database, and candidate binding sites were identified by scanning the hg19 human reference genome using the MATCH-TM algorithm with a matrix similarity score (mSS) threshold of 0.90.

**Neuroblastoma association replication and meta-analysis for rs2168101.** We replicated the association of rs2168101 with neuroblastoma by direct genotyping of rs2168101 in independent Italian (cases = 420, controls = 751) and UK cohorts (cases = 369, controls = 1,109). Meta-analysis across the European-American, African-American, Italian and UK cohorts was performed using the inverse variance method provided in the METAL software program. Beta values (log-odds) and standard errors generated by SNPTEST, as described above, were used as input.

**Survival analysis.** We compared both overall survival and event-free survival over a 10-year follow-up period between G/G versus G/T and T/T rs2168101 genotypes in a case-case comparison between neuroblastoma patients from the European-American cohort. Because rs2168101 genotypes were imputed in this cohort, the most-probable genotype predicted by IMPUTE2 was used for each patient. In the event of insufficient follow-up, all data was right censored. Cox proportional hazard modelling was performed using 20 MDS components to account for population stratification, in addition to MYCN amplification status, as covariates. All statistical analysis and generation of Kaplan-Meier plots was performed in R using the CRAN repository package, 'survival'.

**Total and allele-specific expression analysis.** Total and allele-specific RNA expression analysis was performed based on poly-A-enriched RNA-sequencing data from 127 primary neuroblastoma tumours sequenced through the TARGET project. RNA-seq reads were aligned to the hg19 human reference genome using the STAR aligner (v2.4.0b). Aligned reads were assigned to RefSeq genes using HTSeq (v0.6.1) and normalized to RPKM for total gene expression measurements. DNA genotypes for rs2168101 were obtained either through matched whole-genome sequencing ( $n = 69$ ) or targeted genotyping assays ( $n = 58$  additional tumours). DNA genotypes for rs3750952 were obtained through either matched whole-genome or whole-exome sequencing.

Allele-specific RNA expression analysis was performed from a subset of 45 primary neuroblastoma tumours (out of 127) with the necessary synonymous exonic SNP genotypes (rs3750952 = C/G) to enable measurement of allelic expression by mRNA-seq. As a readout for allelic imbalance of rs3750952, we computed allelic fractions as  $\min(C, G)/(C + G)$ , since phasing between rs3750952 and rs2168101 alleles in each tumour was unknown. Statistical comparison between the two groups was performed by two-sided Welch's  $t$ -test, comparing 12 tumours

heterozygous for rs2168101 (G/T) to the remaining 33 tumours that were homozygous for rs2168101 (G/G) as controls. DNA genotyping for rs2168101 was performed by whole-genome sequencing or a directed genotyping assay, whereas DNA genotyping for rs3750952 was determined from TARGET whole-exome or whole-genome sequencing. Where possible, integrity of sample matching was verified by measurement of genome-wide genotype concordance. All genotypes are reported with respect to the minus strand of the human reference genome, hg19.

To measure allele-specific expression directly at the intronic SNP we first purified the nuclear RNA fraction using the Cytoplasmic and Nuclear RNA purification Kit (Norgen Biotek, 21000) from four neuroblastoma cell lines (SNP rs2168101: SHSY5Y = G/G; NLF = G/T; NGP = G/G; NB1643 = T/T). Ion AmpliSeq Designer v3.4.3 (Life Technologies White Glove service) was used to design amplicons targeting the intronic SNP rs2168101 and three additional exonic SNPs in linkage disequilibrium. Custom AmpliSeq libraries were prepared in triplicate for each cell line, indexed, pooled and sequenced using an Ion 318 Chip on a Personal Genome Machine (Life Technologies). Reads were aligned to the hg19 reference genome and a synthetic genome showing the alternate allele at SNP rs2168101 at hg19 chr11:8255408 to account for any alignment bias. High-quality mapped reads containing the reference G allele or alternative T allele were counted and tested for significant deviation from 50:50 expression using a two-sided one-sample  $t$ -test (null hypothesis that allele fraction = 0.50) across three experimental replicates. Primer pair sequences: rs1042359 forward: 5'-GTGTGGGAGACAAUUTCTTCCUGA-3', reverse: 5'-GCCGGGCGUTACTGAACUT-3'; rs3750952 forward: 5'-CGCAAGAUCAAGGACCGCTAUC-3', reverse: 5'-GATGAGGTUGGCCTTGGTGUA-3'; rs2168101 forward: 5'-CCUTTCUGAAGGAGCGCAAA-3', reverse: 5'-CACTTCCATUAAGGAGATAGCAUCCC-3'; rs204929 forward: 5'-CAAUCTAGGTUAGAGCCGGACAA G-3', reverse: 5'-GTGUCCAGCCGACGCUA-3'.

**Reporter assays.** Primers were designed to clone a 553-bp genomic region (hg19, chr11:8255155-8255707) surrounding the candidate SNP rs2168101 at the GATA transcription factor binding site from neuroblastoma cell lines SKNSH (G/G) and matching site of BE2C (T/T). The cloned region did not contain other statistically significant SNPs at the LMO1 locus. The primers were designed to introduce sequences for restriction sites 5'-XhoI and 3'-BglII, which are present in the MCS of pGL4.26[luc2/minP/Hygro] (Promega, E8441). XhoI/BglII restriction enzyme digested fragments were sequence verified, gel purified, ligated into pGL4.26[luc2/minP/Hygro], transformed into One Shot TOP 10 chemically competent cells (Life Technologies, C4040-10) and grown on LB plates containing 50  $\mu\text{g ml}^{-1}$  ampicillin overnight at 37°C. Colonies positive for the vector containing the insert were grown in 50 ml LB broth containing 50  $\mu\text{g ml}^{-1}$  ampicillin and plasmids were purified using a Qiagen Plasmid Midi Prep Kit (Qiagen, 12143). Transfection into HEK293 cells which were approximately 50% confluent was accomplished using Fugene 6 Transfection reagent (Promega E2691) at a 3  $\mu\text{l}$ :1  $\mu\text{g}$  fugene:DNA ratio. Cells underwent selection in 150  $\mu\text{g ml}^{-1}$  Hygromycin B (Mediatech, 30-240-CR) and individual colonies were picked and grown, and genotypes of constructs were confirmed by fragment size and Sanger sequencing. Subsequently, HEK293 + 553 bp insert cells and HEK294 + vector only cells were grown in 96-well optical plates. On day 2, the cells were transiently fugene transfected with the Renilla expression control vector pGL4.74[hRLuc/TK] (Promega, E6921) at a 1:500 dilution with respect to the luciferase vector. Luciferase assays were carried out 48 h after Renilla transfection using Dual Luciferase Reporter Assay System (Promega, E1910) with read-outs performed on a Dual Injector System for GloMax-Multi Detection System (Promega, E7081). Luciferase expression was normalized to Renilla expression. All reporter assays were performed in quintuplicate (five technical replicates each) across the experimental conditions: (1) HEK293T, (2) HEK293T with empty vector, (3)-(6) four independent clones of HEK293T with T allele construct, and (7)-(10) four independent clones of HEK293T with G allele construct. Results were averaged across technical replicates, normalized to empty vector, and reporter activities for T allele versus G allele clones (four biological replicates each) were analysed by two-sided Welch's  $t$ -test.

Construct replicate allele (G):

GTAGGGGTTGGAGTTCAGCCTGTTTCCCTCCAATGTTGTTCCTCC  
CACATCCTGAGACTTAGGGGTGACCCTGGGTTGAGTGGACTGGTTTA  
TTCTGCTGGGCCAGCGCATGCTGAGTGTGTGCCAGGCGTGCG  
TGTCGGCGCAACATCATCCATGTGAAATATCAGTGTTCATGGGT  
GAGTAGTAATTACTGGGTAATGCTTTAAACCTTTCTCTGAAGGAGCGC  
AAAGCCATTTTCTTAAAGTCAGGATGACATTTAAAGGATTACCATG  
TAGATTTGATTTTATAGATAACACTAAAATGGATCCCAATGGACTTCA  
GCAAAGGGATGCTATCTCTTAATGAAAAGTGCATGCCCCGAGGCTC  
AGGTCCCAGAGCCAGGCTGGGGAAGGAGGAGGGAAGAGGTGTCTG  
CAGGGGGGCGAGGCTGGCAGATTGGGTGGGGGCTAGGTGGGAATGGG  
GAAGGCAGAGCAGGAGGAGGCGCTGGACCTGTGGGGAGCTTATC  
CCTCCATCTGGGAGCAGGAGACATACAGAGCCCTT.

Construct protective allele (T):

GTAGGGGTGGAGTTACGCTGTTTCCCTCCAATGTTGTTCCCCC  
CACATCCTGAGACTAGGGGTGACCTGGGTGAGTGGACTGGTTTA  
TTCTGCTGGGCCCAGCGCATGCATCTGAGTGTGTGCCAGGCGTGCG  
TGTCGGCGCAAACATCATCCATGTGAAATATCAGTGTTCATGGGT  
GAGTAGTAATTACTGGGTAATGCTTTAAACCTTTCTGAAGGAGCGC  
AAAGCCATTTTTTCTAAAGTCAGGAGTACATTAAGGATTACCATG  
TAGATTTGATTTTTATATAACACTAAAATGGATCCCAATGGACTTCAG  
CAAAGGGATGCTATCTCTTAATGAAAAGTGCATGCCCCAGGCTCA  
GGTCCCAGAGCCAGGCTGGGGAAGGAGGGAGGGAAGAGGTGTCTGC  
AGGGGGCAGGCTGGCAGATTGGGTGGGGGGCTAGGTGGGAATGGG  
GAAGGCAGAGCAGGAGGGAGGGCCTGGACCCTGTGGGGAGCTTATC  
CCTCCATCTGGGGAGCAGGAGACTACAGAGCCCCCT.

**Cell culture and protein lysates.** Jurkat T-ALL and neuroblastoma cell lines were sourced from the American Type Tissue Culture Collection, and kept in growth medium of RPMI + 10% heat-inactivated FCS with 1% penicillin-streptomycin, as previously described<sup>13</sup>. Cells were lysed for protein, with subsequent protein quantified by spectrophotometry, as previously described<sup>21</sup>. Protein was resolved on 8–14% Tris-Bis gels, transferred to PVDF membranes, blocked and subjected to primary and secondary antibodies, as previously described<sup>21</sup>. Primary antibodies were anti-GATA3 (Pierce Biotechnology, 1:1,000), anti-LMO1 (Bethyl Laboratories, 1:1,000) and  $\alpha$ -tubulin (Cell Signaling Technologies, 1:1,000). Blots were developed with secondary horseradish peroxidase (HRP)-conjugated antibodies (Cell Signaling Technologies, 1:5,000) and Protein-plus Dura ECL Reagent (Thermo-Fisher Scientific). All cell lines are genotyped semiannually to assure identity and also tested routinely for mycoplasma contamination.

**Genome-wide occupancy analysis.** ChIP coupled with massively parallel DNA sequencing (ChIP-seq) was performed as previously described<sup>22,23</sup>. The following antibodies were used for ChIP: anti-H3K27ac (Abcam, ab4729) and anti-GATA3 (Santa Cruz, sc-22206X). For each ChIP, 10  $\mu$ g of antibody was added to 3 ml of sonicated nuclear extract. Illumina sequencing, library construction and ChIP-seq analysis methods were previously described<sup>23</sup>.

**ChIP-seq processing.** Reads were aligned to build hg19 of the human genome using bowtie with parameters -k 2 -m 2 -e 70 -best and -l set to the read length<sup>24</sup>. For visualization in the UCSC genome browser in Fig. 3a, c and Extended Data Fig. 6 (ref. 25), WIG files were created from aligned ChIP-seq read positions using MACS 1.4.2 with parameters -w -S -space = 50 -nomodel -shiftsize = 200 to artificially extend reads to be 200bp and to calculate their density in 50-bp bins<sup>26</sup>. Read counts in 50-bp bins were then normalized to the millions of mapped reads, giving reads per million values.

**ChIP-seq allele specificity analysis.** To determine preferential ChIP-seq coverage of one allele, which implies preferential binding of protein to one allele versus another, we counted the reads at rs2168101 using samtools mpileup<sup>27</sup>. By using the aligned reads described above, this gave us a count of reads with a given base at this position. The fraction of reads with the risk allele versus the protective allele is reported in Fig. 3b. Statistical tests for preferential allelic binding were performed by two-sided binomial test.

**Enriched regions.** Regions enriched in ChIP-Seq signal were identified twice using MACS with corresponding control and parameters -keep-dup = all and -p 1e-9 or -keep-dup = 1 and -p 1e-9. Super-enhancers in SHSY5Y and KELLY were identified using ROSE (https://bitbucket.org/young\_computation/rose)<sup>18,28</sup> with modifications based on ref. 14. In brief, peaks of H3K27ac were identified using MACS as described above and their union was used as constituent enhancers. These peaks were stitched computationally if they were within 12,500 bp of each other, although peaks fully contained within  $\pm 2,000$  bp from a RefSeq promoter were excluded from stitching. These stitched enhancers were ranked by their H3K27ac signal (length  $\times$  density) with input signal in the corresponding region subtracted. Super-enhancers were separated from typical enhancers by geometrically determining the point at which the line  $y = x$  is tangent to the curve of stitched enhancer rank versus stitched enhancer signal. Those stitched enhancers above this point are considered super-enhancers.

To account for the known focal amplification of the MYCN locus in KELLY, BE2, BE2C and NGP neuroblastoma cells, which contain enhancers, we modified our pipeline slightly. Because MACS is insensitive for the identification of peaks in focally amplified DNA, we identified peaks of H3K27ac versus input using MACS2 callpeak (https://pypi.python.org/pypi/MACS2) with parameters -broad -keep-dup = 1 -p 1e-9 and -broad -keep-dup = all -p 1e-9. The union of these MACS2 calls was used as constituent enhancers for ROSE with the remaining parameters as described above. For Fig. 3d, most of the curve represents the

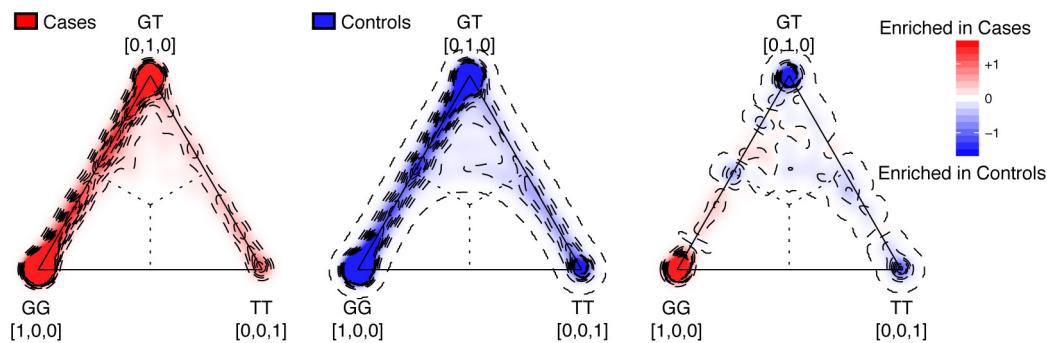
analysis performed using MACS-identified constituents; the rank and signal of the MYCN-associated enhancer comes from this MACS2-identified set of constituents to remain consistent with the conclusions and methods as previously described<sup>14</sup>. The curve output from the MACS-identified enhancers was vertically compressed and a point representing the signal of the MYCN-associated super-enhancer from the MACS2-identified enhancers was added in Illustrator. Super-enhancers were assigned to the single expressed RefSeq transcript whose transcription start site was nearest the centre of the stitched region. Expressed genes were in the top 2/3 of RefSeq transcripts ranked by their promoter (transcription start site  $\pm 500$  bp) H3K27ac signal determined by bamToGFF (https://github.com/BradnerLab/pipeline) with parameters -e 200 -m 1 -r -d.

**Clone cell generation.** LMO1 cDNA was amplified from pcDNA3-LMO1 and subcloned into the XhoI and NotI site of the lentiviral vector pOZ-FHN. Lentivirus expressing FH-LMO1 was propagated in HEK293T cells by cotransfection with psPAX2 and pMD2.G plasmids (adgene) using FUGENE 6 (Roche) by standard methodologies<sup>29</sup>. Viral supernatant was recovered and KELLY cells were infected with lentivirus expressing FH-LMO1 or empty vector alone, as previously described<sup>13</sup>. Cells were sorted for expression of the IL2R, and positive expression was used to establish single cell clones. Expression of FH-LMO1 was assessed by western blotting as above to confirm overexpression.

**siRNA and growth assays.** SHSY5Y, KELLY and KELLY clone cells were reverse transfected with 100 nM concentrations of either non-targeted (control siRNA-1) or GATA3-targeted siRNA-1 or -2 (Ambion) for 6 h with lipofectamine 2000 (1:1,000) in OptiMax I before being replated into growth assays in normal RPMI growth media. Cells ( $2 \times 10^5$ ) were replated in triplicate for counting at 24, 48 and 72 h post-transfection by manual hemocytometry. Cells ( $5 \times 10^5$ ) were replated for protein lysates at the same time points. All experiments were repeated in triplicate, with a technical replicate number of 9 for all cell growth assays as described<sup>30</sup>. Statistical tests were performed by two-sided Welch's *t*-test.

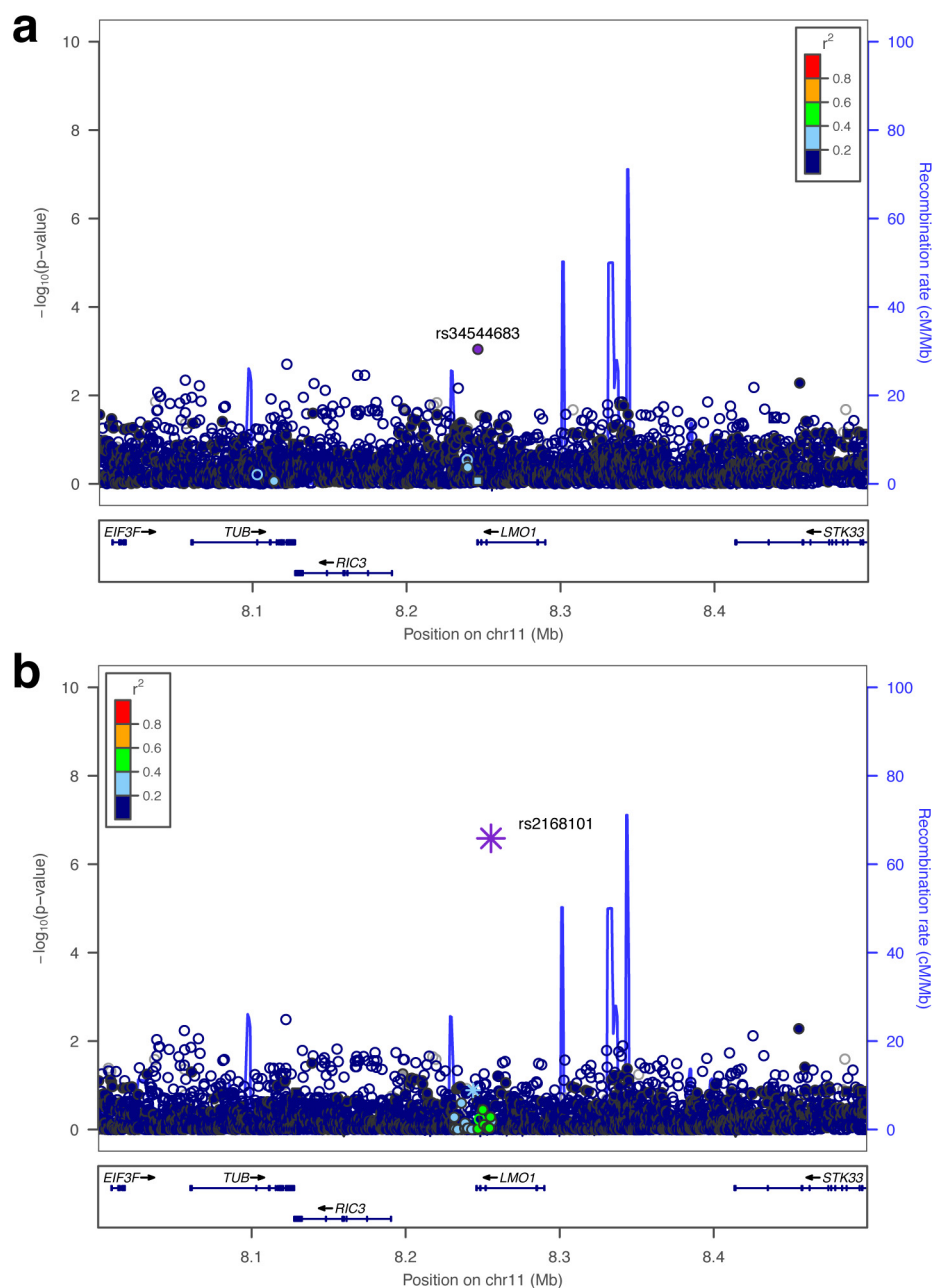
**Data access.** GWAS and sequencing data used for this analysis are available in dbGaP under accession phs000124 and phs000467. The tumour genomics data are also available through the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) data matrix portal (http://target.nci.nih.gov/dataMatrix/TARGET\_DataMatrix.html). Data generated through the ENCODE project including DNase I hypersensitivity sequencing and ChIP-sequencing data were obtained from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/. Aligned sequencing read (bam) files were used as provided from the FTP site. The mammalian evolutionary conservation track representing 46 mammalian species (phastCons46wayPlacental) was obtained from the UCSC Table Browser http://genome.ucsc.edu/cgi-bin/hgTables?command=start. JASPAR-annotated transcription factor binding site position frequency matrices were obtained from http://jaspar.genereg.net/html/DOWNLOAD/JASPAR\_CORE/pfm/nonredundant/pfm\_all.txt. New ChIP-seq data sets generated in this study are available under super series GSE65664.

20. Latorre, V. *et al.* Replication of neuroblastoma SNP association at the *BARD1* locus in African-Americans. *Cancer Epidemiol. Biomarkers Prev.* **21**, 658–663 (2012).
21. Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
22. Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols* **1**, 729–748 (2006).
23. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521–533 (2008).
24. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
25. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
26. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
28. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
29. Nakatani, Y. & Ogryzko, V. Immunoaffinity purification of mammalian protein complexes. *Methods Enzymol.* **370**, 430–444 (2003).
30. Durbin, A. D. *et al.* JNK1 determines the oncogenic or tumor-suppressive activity of the integrin-linked kinase in human rhabdomyosarcoma. *J. Clin. Invest.* **119**, 1558–1570 (2009).



**Extended Data Figure 1 | The imputed SNP, rs2168101, is associated with neuroblastoma, and the risk 'G' allele is enriched in neuroblastoma cases.** Ternary density plots of genotype probability vectors [P(G/G), P(G/T), P(T/T)] output from IMPUTE2 for rs2168101 in the European-American cohort. Vertices represent 'perfect' confidence calls in which P(genotype) = 1; dotted lines represent decision boundaries for genotype calling based on most probable genotype. All plots were normalized by

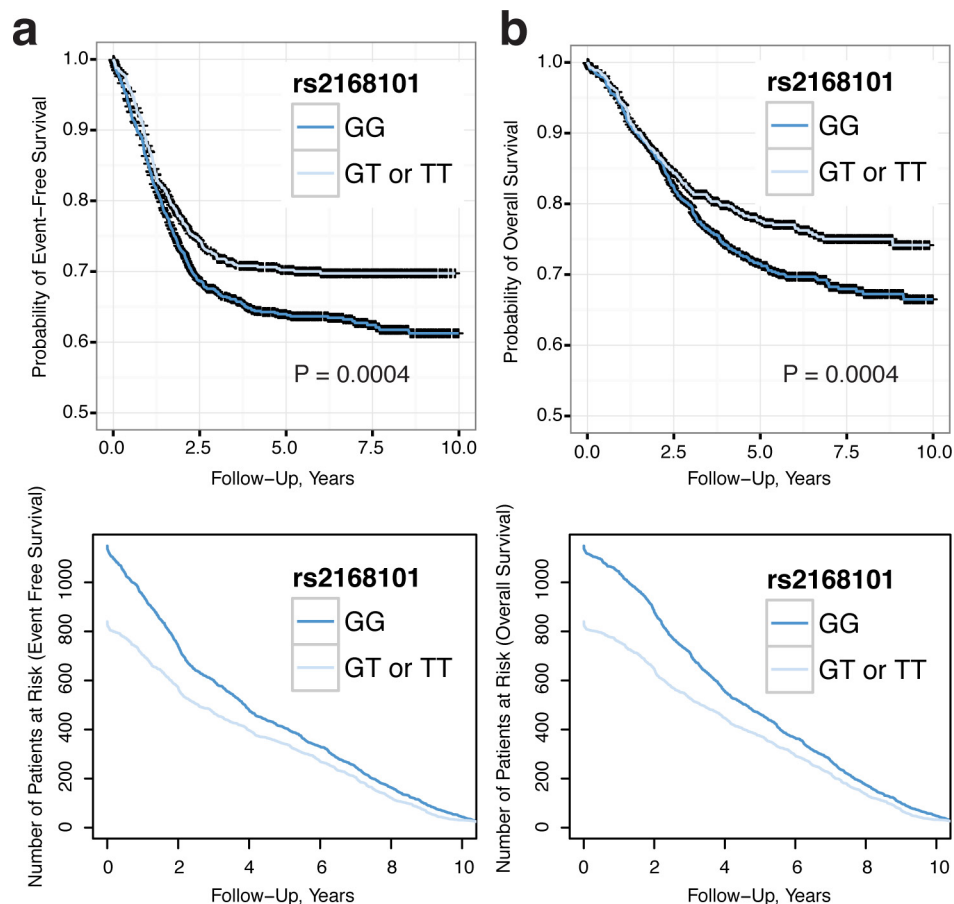
the total number of individuals studied and subjected to 2D Gaussian kernel smoothing. Left, 2,101 cases (red); centre, 4,202 controls (blue); right, difference between cases and controls highlights enrichment of G/G genotype (homozygous risk) in cases and of G/T and T/T genotypes in controls. Validation efforts using PCR-based genotyping in 146 out of 2,101 European-American cases confirmed an 86% concordance with imputation based on most probable genotypes (Supplementary Table 1).



**Extended Data Figure 2 | Conditional analysis reveals a single neuroblastoma association signal at the *LMO1* locus and that rs2168101 is the most associated variant.** **a**, Imputation-based neuroblastoma association study conditional on rs2168101. No variants remain significant after conditioning on rs2168101 (most significant variant: rs34544683, nominal  $P = 9.0 \times 10^{-4}$ , Bonferroni  $P = 1$ ). **b**, Reciprocal analysis conditioned on each of 27 SNPs with a nominal  $P < 1 \times 10^{-5}$ . For rs2168101, the maximum (least significant)  $P$  value across all

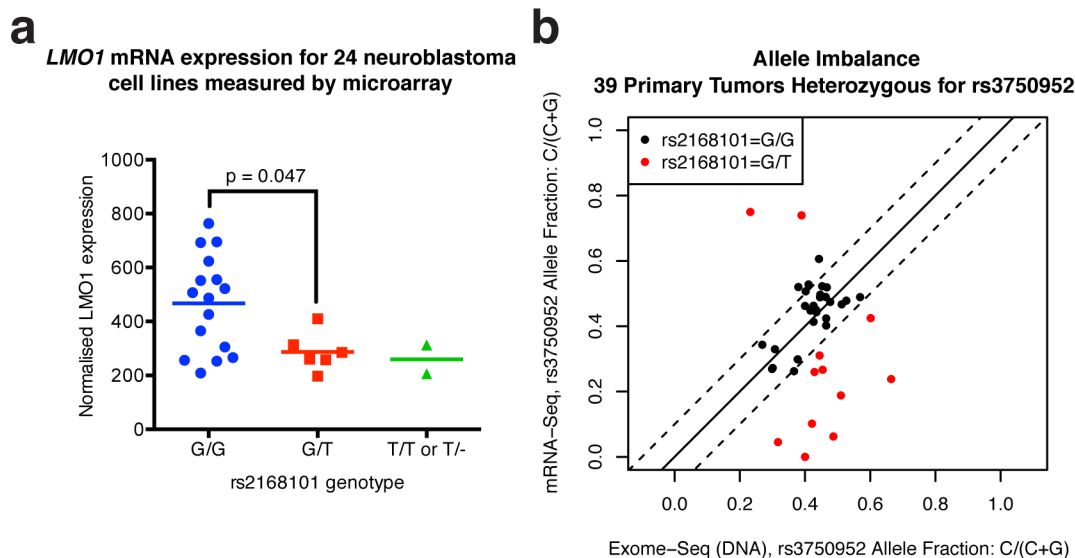
non-rs2168101 conditional tests is shown, in order to illustrate the extent to which the signal at rs2168101 can be accounted for by other variants (a similar maximum  $P$  value statistic is plotted for other variants). Notably, rs2168101 remained significant (worst-case nominal  $P = 2.6 \times 10^{-7}$ , Bonferroni  $P = 0.002$ ) across all tests. These results are consistent with a single underlying signal at the *LMO1* locus, and re-affirm that rs2168101 is the single best causal SNP candidate because its association with neuroblastoma cannot be accounted for by other single variants.





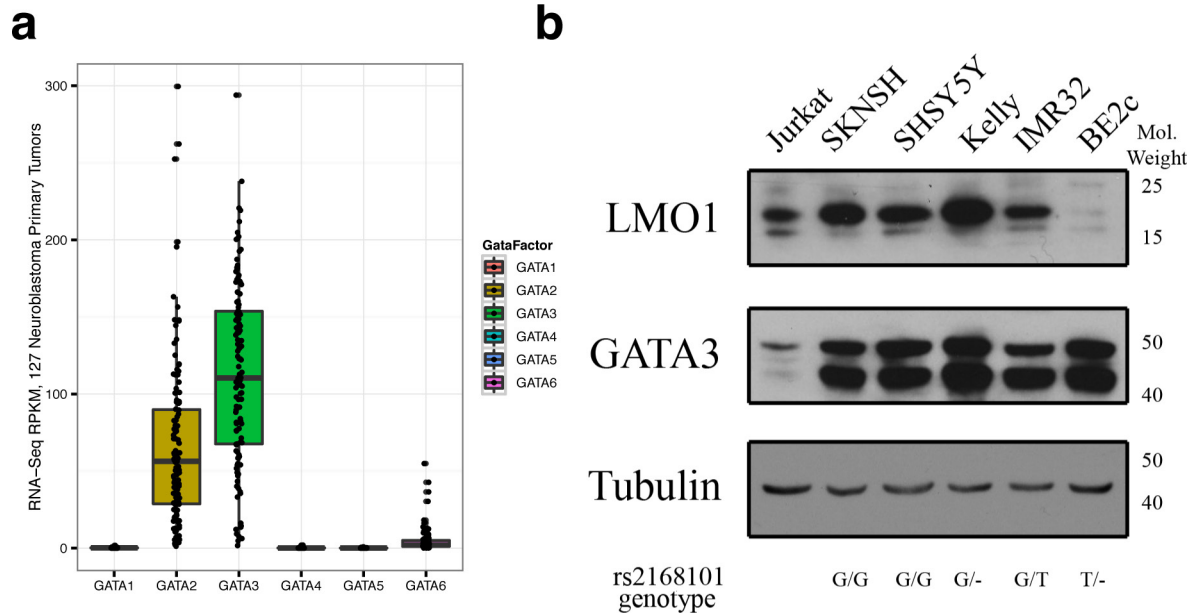
**Extended Data Figure 3 | The risk G allele of rs2168101 is associated with decreased event-free and overall survival in the European-American discovery cohort.** Because genotypes for rs2168101 are imputed within the European-American discovery cohort, the most likely genotype for each neuroblastoma case was called based on the maximum of P(G/G), P(G/T) and P(T/T) from IMPUTE2. *P* values reflect Cox proportional hazards regressions adjusted for *MYCN* amplification status and the first 20 MDS components to adjust for population stratification.

**a**, Kaplan–Meier plot for event-free survival. Neuroblastoma cases with rs2168101 = G/G versus rs2168101 = G/T or T/T showed significantly worse event-free survival ( $P = 0.0004$ ). **b**, Kaplan–Meier plot for overall survival. Neuroblastoma cases with rs2168101 = G/G versus rs2168101 = G/T or T/T showed significantly worse overall survival ( $P = 0.0004$ ). Censored data points are shown as black crosses. Number of at risk patients at every time point for both event-free survival and overall survival are plotted below each respective Kaplan–Meier plot.



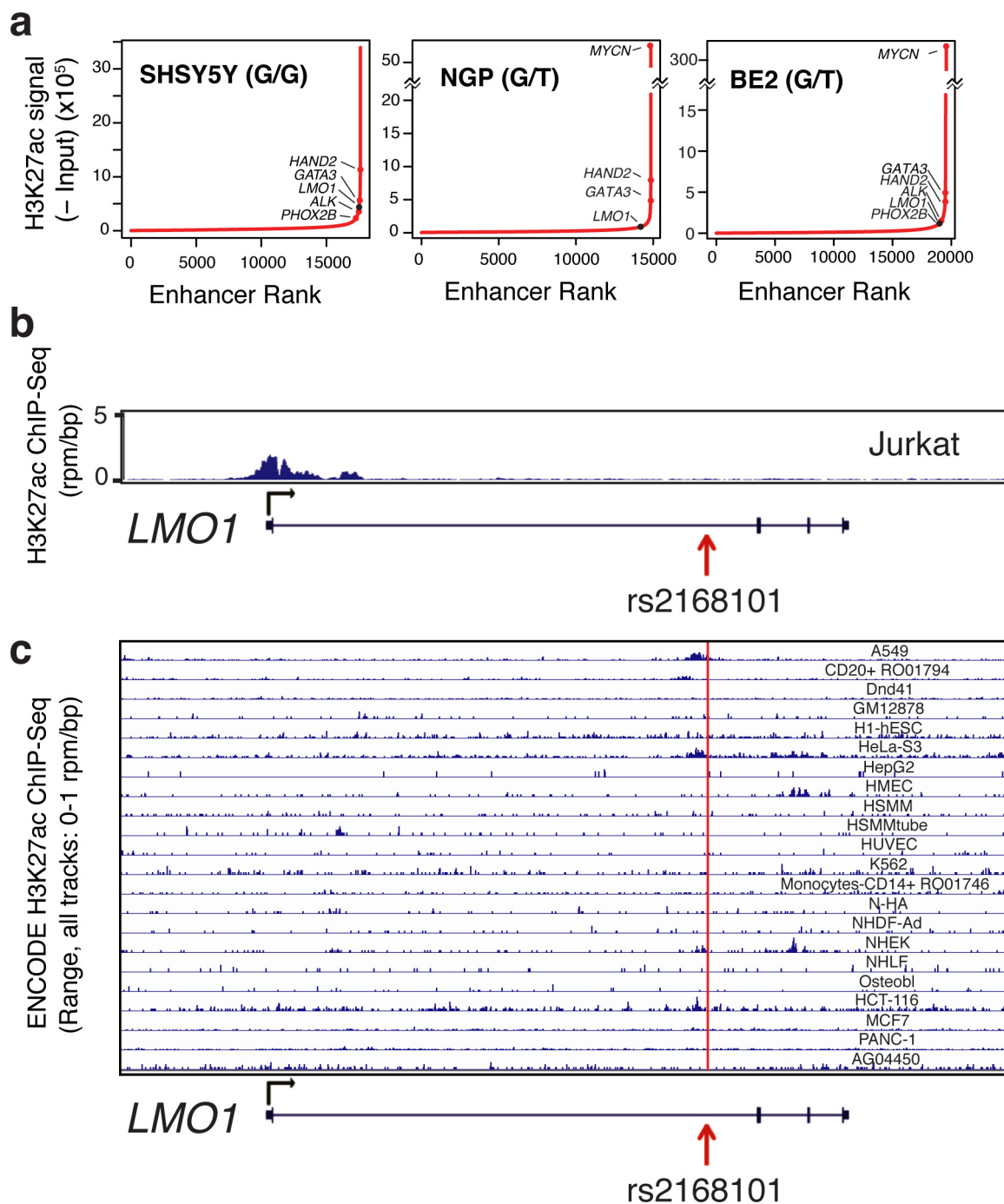
**Extended Data Figure 4 | rs2168101 genotype is associated with total and allele-specific *LMO1* expression in neuroblastoma cell lines and primary tumours, and allele-specific expression differences are not driven by somatic DNA copy number alterations.** **a**, Neuroblastoma cell line *LMO1* mRNA expression as quantified by Affymetrix U95Av2 oligonucleotide arrays and normalized as described<sup>11</sup> was significantly higher in cell lines harbouring homozygous risk alleles (G/G) compared to heterozygous alleles (G/T) ( $P = 0.047$ , Mann–Whitney two-tailed). **b**, Allele-specific expression measured by RNA-seq from primary neuroblastoma tumours. Since rs2168101 is an intronic SNP that is spliced out in mRNA, the synonymous exonic SNP rs3750952 was used as a surrogate for measuring allele-specific expression in 39 primary tumours

which are heterozygous for rs3750952 (C/G genotype). The DNA allelic fraction for rs3750952 determined by whole-exome sequencing is plotted on the *x* axis, whereas the RNA allele fraction for rs3750952 determined by mRNA-seq is plotted on the *y* axis. The solid line indicates where DNA and RNA allele fractions are equal and dotted lines indicate the boundary where DNA and RNA allele fractions are within 10% of each other. Tumours that are heterozygous for rs2168101 (G/T genotype, red dots) exhibit greater RNA allelic imbalance ( $P = 5.3 \times 10^{-5}$ ) than homozygous controls (rs2168101 = G/G genotype, black dots). By contrast, DNA allelic imbalance is no different between G/T versus G/G tumours ( $P = 0.79$ ), indicating that a *cis*-acting regulatory mechanism, rather than somatic DNA alterations, drives *LMO1* allelic expression differences.



**Extended Data Figure 5 | Expression of LMO1 and GATA-family transcription factors in neuroblastoma primary tumours and cell lines.** **a**, RPKM expression measurements from mRNA-seq are summarized via boxplots for 127 primary neuroblastoma tumours for paralogues *GATA1* through *GATA6*. Both *GATA2* (median RPKM: 56) and *GATA3* (median RPKM: 110) are more highly expressed by 1–4 orders of

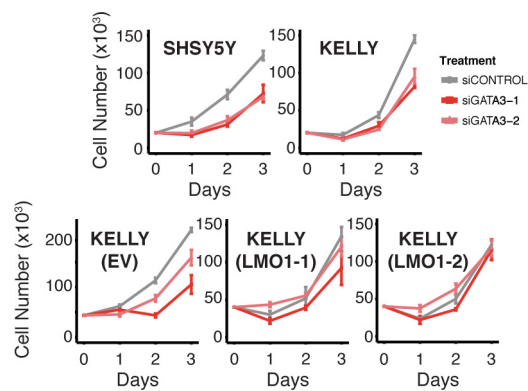
magnitude on average compared to other members of the GATA family in neuroblastoma. **b**, Neuroblastoma cell lines were lysed for protein and resolved by SDS-PAGE as previously described<sup>21</sup>. Jurkat T-ALL cells are shown as a positive control for LMO1 and GATA3 expression. Data are representative of at least three independent blots. The rs2168101 genotype is shown below individual cell lines.



**Extended Data Figure 6 | The *LMO1* super-enhancer is observed in neuroblastoma cell lines containing the G allele of rs2168101 and is highly tissue-specific.** **a**, H3K27ac signal across all enhancers in SHSY5Y (*MYCN* not amplified; rs2168101 = G/G), BE2 (*MYCN*-amplified; rs2168101 = G/T) and NGP (*MYCN*-amplified; rs2168101 = G/T) is shown. Enhancers are ranked by their signal of H3K27ac minus input signal and are geometrically divided into two populations (see Methods). Super-enhancers are those at the high end of the population and are associated with key genes in neuroblastoma, highlighted on the curve.

*LMO1*-associated super-enhancers were identified in BE2, KELLY and SHSY5Y cells, which all contain the G allele of rs2168101, but not in BE2C cells in which the G allele is absent. **b**, H3K27ac ChIP-seq in the Jurkat cell line. **c**, All ENCODE non-neuroblastoma cell lines with H3K27ac ChIP-seq profiling. All non-neuroblastoma cell lines considered showed little to no evidence for an active enhancer element within the first intron of the *LMO1* gene locus, consistent with a tissue and disease-specific enhancer overlying the neuroblastoma causal SNP rs2168101.





**Extended Data Figure 7 | Depletion of GATA3 results in suppression of cell growth that is rescued by forced LMO1 expression in neuroblastoma.** Neuroblastoma cells SHSY5Y, KELLY, KELLY overexpressing control vector (EV) and KELLY with forced LMO1 overexpression (LMO1-1 and LMO1-2) were treated with non-targeted (siControl) or GATA3-targeting (siGATA3-1, siGATA3-2) siRNAs and cells were counted at 24, 48 and 72 h after transfection. Rescue of suppressed cell growth after GATA3 depletion by forced LMO1 expression in LMO1-1 and LMO1-2 after 72 h is shown on the bottom. Growth curves over the time of 72 h are shown (to accompany Fig. 3f). Error bars denote  $\pm$ s.e.m.,  $n = 9$  technical replicates.

**Extended Data Table 1 | Germline variants from 1000 Genomes Project with  $P < 1 \times 10^{-5}$  association with neuroblastoma susceptibility from imputation-based SNPTEST analysis of European-American cohort**

Variant ID (rsID)	Chromosome	Position (hg19)	Alleles (Ref/Alt)	Alt Allele Frequency Cases	Alt Allele Frequency Controls	P-Value <sup>†</sup>	Odds Ratio <sup>†</sup>
rs191871553	11	8222464	C/T	0.035 (n=2101)	0.054 (n=4202)	7.49E-06	0.64 (0.53-0.78)
rs11041809	11	8231605	A/G	0.498 (n=2101)	0.440 (n=4202)	1.13E-07	0.80 (0.74-0.87)
rs11041811	11	8231665	C/T	0.492 (n=2101)	0.434 (n=4202)	1.28E-07	0.80 (0.74-0.87)
rs11041812	11	8231684	C/T	0.492 (n=2101)	0.433 (n=4202)	1.22E-07	0.80 (0.74-0.87)
rs11041813	11	8235207	T/C	0.478 (n=2101)	0.420 (n=4202)	1.67E-07	0.81 (0.75-0.87)
rs10839999	11	8236083	G/A	0.480 (n=2101)	0.423 (n=4202)	5.06E-07	0.81 (0.75-0.88)
rs10769885	11	8236262	C/A	0.513 (n=2101)	0.453 (n=4202)	3.77E-08	0.80 (0.74-0.87)
rs4758049	11	8238428	A/C	0.511 (n=2101)	0.452 (n=4202)	7.48E-08	0.81 (0.74-0.87)
rs4758050	11	8238545	G/C	0.511 (n=2101)	0.452 (n=4202)	7.34E-08	0.81 (0.74-0.87)
rs4758051	11	8238639	G/A	0.510 (n=2101)	0.452 (n=4202)	1.22E-07	0.81 (0.75-0.87)
rs10840000	11	8240113	G/C	0.509 (n=2101)	0.450 (n=4202)	6.22E-08	0.80 (0.74-0.87)
rs7933766	11	8240464	G/A	0.511 (n=2101)	0.453 (n=4202)	2.09E-07	0.81 (0.75-0.88)
rs11041816	11	8243798	A/G	0.397 (n=2101)	0.456 (n=4202)	8.99E-10	0.77 (0.71-0.84)
rs4315061	11	8247020	T/C	0.425 (n=2101)	0.490 (n=4202)	1.25E-09	0.78 (0.72-0.84)
rs72474792	11	8247885	TATAAAA/T	0.524 (n=2101)	0.456 (n=4202)	2.04E-10	0.77 (0.71-0.84)
rs12797723	11	8247984	C/T	0.443 (n=2101)	0.514 (n=4202)	2.05E-10	0.77 (0.71-0.84)
rs2290451	11	8248440	C/G	0.295 (n=2101)	0.255 (n=4202)	8.20E-06	1.23 (1.12-1.34)
rs7952320	11	8250143	G/C	0.408 (n=2101)	0.480 (n=4202)	3.03E-11	1.31 (1.21-1.42)
rs4758317	11	8250811	C/A	0.514 (n=2101)	0.447 (n=4202)	5.76E-10	0.78 (0.72-0.84)
rs11041820	11	8251438	G/A	0.294 (n=2101)	0.253 (n=4202)	6.77E-06	1.23 (1.12-1.34)
rs3750952	11	8251921	G/C	0.408 (n=2101)	0.481 (n=4202)	1.89E-11	0.76 (0.70-0.83)
rs110419	11	8252853	A/G	0.441 (n=2101)	0.511 (n=4202)	3.16E-10	0.78 (0.72-0.84)
rs110420	11	8253049	T/C	0.441 (n=2101)	0.511 (n=4202)	3.36E-10	0.78 (0.72-0.84)
rs204928	11	8254433	A/G	0.444 (n=2101)	0.512 (n=4202)	9.85E-10	0.78 (0.72-0.85)
rs204926	11	8255106	G/A	0.440 (n=2101)	0.510 (n=4202)	1.97E-11	0.76 (0.70-0.82)
rs2168101	11	8255408	C/A	0.242 (n=2101)	0.313 (n=4202)	4.14E-16	0.67 (0.61-0.74)
rs7948497	11	8255855	C/G	0.479 (n=2101)	0.419 (n=4202)	4.05E-10	1.30 (1.20-1.41)

\*Forward strand hg19, imputed genotypes from IMPUTE2, frequencies as reported by SNPTEST.

†SNPTEST, frequentist score test with additive model, adjusted for gender and top 20 MDS components.

Extended Data Table 2 | Clinical characteristics for patients in referenced sequencing data sets

Characteristic	Whole Exome Sequencing (Blood and Tumor) N=222	Whole Genome Sequencing (Blood and Tumor) N = 136	<i>LMO1</i> Targeted Sequencing (Blood) N = 183	Whole Transcriptome mRNA Sequencing (Tumor) N = 127
<b>Age</b>				
< 18 mos	0 (0%)	32 (24%)	82 (45%)	8 (6%)
>= 18 mos	219 (100%)	103 (76%)	101 (55%)	119 (94%)
Not Available	3	1	0	0
<b>INSS Stage<sup>†</sup></b>				
Stage 1	0 (0%)	0 (0%)	39 (21%)	0 (0%)
Stage 2A	0 (0%)	0 (0%)	13 (7%)	0 (0%)
Stage 2B	0 (0%)	1 (1%)	18 (10%)	0 (0%)
Stage 3	0 (0%)	6 (4%)	27 (15%)	6 (5%)
Stage 4	219 (100%)	105 (78%)	78 (43%)	121 (95%)
Stage 4S	0 (0%)	23 (17%)	8 (4%)	0 (0%)
Not Available	3	1	0	0
<b>MYCN</b>				
Not Amplified	143 (67%)	102 (76%)	151 (83%)	95 (75%)
Amplified	71 (33%)	32 (24%)	30 (17%)	31 (25%)
Not Available	8	2	2	1
<b>Histology</b>				
Favorable	4 (2%)	29 (23%)	95 (54%)	9 (8%)
Unfavorable	187 (98%)	96 (77%)	82 (46%)	107 (92%)
Not Available	31	11	6	11
<b>DNA Index</b>				
Hyperdiploid	117 (54%)	81 (61%)	121 (67%)	67 (53%)
Diploid	98 (46%)	52 (39%)	59 (33%)	59 (47%)
Not Available	7	3	3	1
<b>Risk</b>				
Low	0 (0%)	15 (11%)	64 (35%)	0 (0%)
Intermediate	0 (0%)	14 (10%)	49 (27%)	6 (5%)
High	219 (100%)	106 (79%)	69 (38%)	121 (95%)
Not Available	3	1	1	0

\*There is an overlap of 59 neuroblastoma patients with both whole-exome and whole-genome sequencing. Patients with targeted sequencing are all unique and do not overlap with whole-exome or whole-genome cases, yielding 482 unique patients with exonic DNA sequencing of *LMO1*.

†International Neuroblastoma Staging System (INSS).

Extended Data Table 3 | Association of rs2168101 with clinical/biological co-variables

Clinical/Biological Co-variate	rs2168101 genotypes <sup>*</sup>			Association result	
	GG	GT	TT	P-Value <sup>†</sup>	Odds Ratio <sup>‡</sup>
Stage <sup>‡</sup> 4	530 (62%)	280 (33%)	49 (6%)	0.01198	0.81 (0.69-0.95)
Not Stage 4	611 (56%)	400 (37%)	74 (7%)		
MYCN Amplified	183 (55%)	115 (34%)	36 (11%)	0.00297	1.39 (1.12-1.73)
MYCN Non-Amplified	881 (59%)	525 (35%)	83 (6%)		
High-Risk	523 (63%)	263 (32%)	47 (6%)	0.00174	0.76 (0.65-0.90)
Not High-Risk	594 (56%)	398 (37%)	73 (7%)		
Unfavorable Histology	454 (61%)	237 (32%)	48 (6%)	0.14479	0.88 (0.73-1.05)
Favorable Histology	527 (57%)	336 (36%)	62 (7%)		
DNA Index Hyperdiploid	685 (59%)	412 (35%)	71 (6%)	0.32009	0.91 (0.76-1.09)
DNA Index Diploid	324 (57%)	198 (35%)	43 (8%)		
Age $\geq$ 18 mos	621 (61%)	346 (34%)	55 (5%)	0.01448	0.82 (0.69-0.96)
Age < 18 mos	529 (57%)	338 (36%)	68 (7%)		

\*Reverse strand hg19, imputed genotypes from IMPUTE2, genotype frequencies as reported by SNPTEST.

†SNPTEST, frequentist score test with additive model, adjusted for gender and top 20 MDS components.

‡International Neuroblastoma Staging System.



# A mechanism for the suppression of homologous recombination in G1 cells

Alexandre Orthwein<sup>1\*</sup>, Sylvie M. Noordermeer<sup>1\*</sup>, Marcus D. Wilson<sup>1</sup>, Sébastien Landry<sup>1</sup>, Radoslav I. Enchev<sup>2</sup>, Alana Sherker<sup>1,3</sup>, Megan Munro<sup>1</sup>, Jordan Pinder<sup>4</sup>, Jayme Salsman<sup>4</sup>, Graham Dellaire<sup>4</sup>, Bing Xia<sup>5</sup>, Matthias Peter<sup>2</sup> & Daniel Durocher<sup>1,3</sup>

**DNA repair by homologous recombination<sup>1</sup> is highly suppressed in G1 cells<sup>2,3</sup> to ensure that mitotic recombination occurs solely between sister chromatids<sup>4</sup>. Although many homologous recombination factors are cell-cycle regulated, the identity of the events that are both necessary and sufficient to suppress recombination in G1 cells is unknown. Here we report that the cell cycle controls the interaction of BRCA1 with PALB2–BRCA2 to constrain BRCA2 function to the S/G2 phases in human cells. We found that the BRCA1–interaction site on PALB2 is targeted by an E3 ubiquitin ligase composed of KEAP1, a PALB2-interacting protein<sup>5</sup>, in complex with cullin-3 (CUL3)–RBX1 (ref. 6). PALB2 ubiquitylation suppresses its interaction with BRCA1 and is counteracted by the deubiquitylase USP11, which is itself under cell cycle control. Restoration of the BRCA1–PALB2 interaction combined with the activation of DNA-end resection is sufficient to induce homologous recombination in G1, as measured by RAD51 recruitment, unscheduled DNA synthesis and a CRISPR–Cas9-based gene-targeting assay. We conclude that the mechanism prohibiting homologous recombination in G1 minimally consists of the suppression of DNA-end resection coupled with a multi-step block of the recruitment of BRCA2 to DNA damage sites that involves the inhibition of BRCA1–PALB2–BRCA2 complex assembly. We speculate that the ability to induce homologous recombination in G1 cells with defined factors could spur the development of gene-targeting applications in non-dividing cells.**

The breast and ovarian tumour suppressors BRCA1, PALB2 and BRCA2 promote DNA double-strand break (DSB) repair by homologous recombination<sup>7–9</sup>. BRCA1 promotes DNA-end resection to produce the single-stranded (ss)DNA necessary for homology search and strand invasion<sup>1</sup> and it also interacts with PALB2 (refs 10–12) to direct the recruitment of BRCA2 (ref. 10) and RAD51 (refs 13 and 14) to DSB sites. The accumulation of BRCA1 on the chromatin that flanks DSB sites is suppressed in G1 cells<sup>15</sup>, reminiscent of the potent inhibition of homologous recombination in this phase of the cell cycle. Since the inhibition of BRCA1 recruitment in G1 is dependent on the 53BP1 and RIF1 proteins<sup>15,16</sup>, two inhibitors of end resection<sup>15–19</sup>, this regulation of BRCA1 was originally viewed in light of its function in DNA-end processing.

However, as BRCA1 is also involved in promoting the recruitment of BRCA2 through its interaction with PALB2, we asked whether inducing BRCA1 recruitment to DSB sites in G1, through mutation of 53BP1 (also known as *TP53BP1*) by genome editing (53BP1Δ; Extended Data Fig. 1a–c) also resulted in BRCA2 accumulation into ionizing-radiation-induced foci. To our surprise, and in contrast with BRCA1, we found that neither BRCA2 nor PALB2 are recruited to G1 DSB sites in U2OS cells lacking 53BP1 at ionizing radiation doses ranging from 2 to 20 Gy (Fig. 1a, b and Extended Data Fig. 1d, e).

Since BRCA1 and PALB2 interact directly<sup>10,11</sup>, this result suggested that G1 cells may block BRCA2 recruitment by suppressing the BRCA1–PALB2 interaction. Indeed, while PALB2 interacts with BRCA2 irrespective of cell cycle position, it interacts efficiently with BRCA1 only during S phase (Fig. 1c). The presence of DNA damage led to the loss of the residual PALB2–BRCA1 interaction in G1 whereas it had little impact on the assembly of the BRCA1–PALB2–BRCA2 complex in S phase (Fig. 1c). Since all proteins were expressed in G1 (Fig. 1c), our results suggest that the assembly of the BRCA1–PALB2–BRCA2 complex is controlled during the cell cycle, possibly to restrict the accumulation of BRCA2 at DSB sites to the S/G2 phases.

We confirmed these results using a single-cell assay assessing the colocalization, at an integrated *lacO* array<sup>20</sup>, of an mCherry-tagged LacR–BRCA1 fusion protein with GFP-tagged PALB2 (Extended Data Fig. 2a). This LacR/*lacO* system recapitulated the cell-cycle-dependent and DNA-damage-sensitive BRCA1–PALB2 interaction (Extended Data Fig. 2b) and enabled us to determine that sequences on PALB2, located outside its amino-terminal BRCA1–interaction domain (residues 1–50) were responsible for the cell-cycle-dependent regulation of its association with BRCA1 (Extended Data Fig. 2c, d). Further deletion mutagenesis identified a single region, encompassed within residues 46–103 in PALB2 (Extended Data Fig. 2e, f) responsible for the cell-cycle-dependent regulation of the BRCA1–PALB2 interaction. This region corresponds to the interaction site for KEAP1 (ref. 5), identifying this protein as a candidate regulator of the BRCA1–PALB2 interaction.

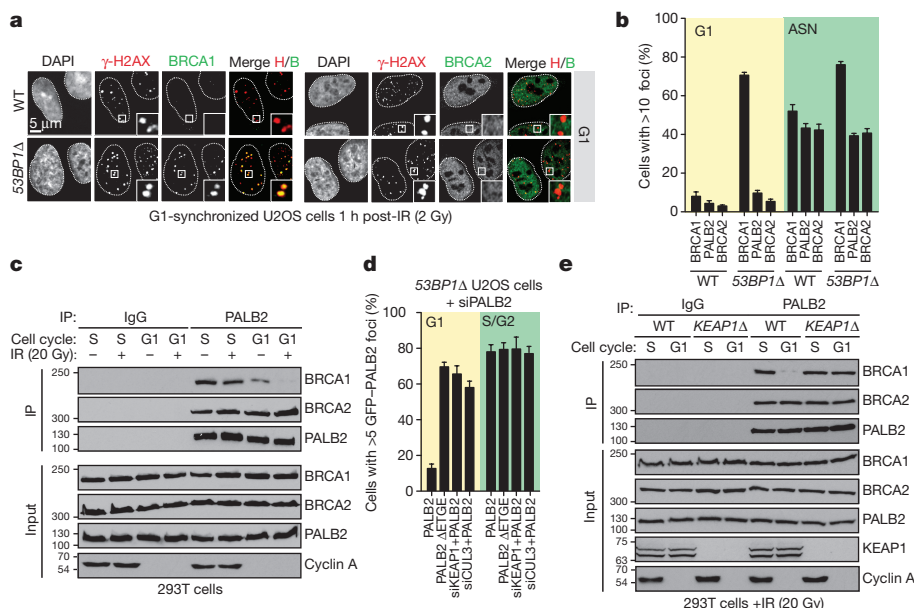
KEAP1 is a substrate adaptor for a CUL3–RING ubiquitin (Ub) ligase (CRL3) that targets the antioxidant regulator NRF2 for proteasomal degradation<sup>21</sup> and recognizes an ‘ETGE’ motif on both PALB2 and NRF2 through its KELCH domain<sup>5</sup>. Depletion of KEAP1 from 53BP1Δ cells, or deletion of the ETGE motif in full-length PALB2 (PALB2 ΔETGE) induced PALB2 ionizing-radiation-induced focus formation in G1 cells (Fig. 1d and Extended Data Fig. 3a). Furthermore, in cells in which KEAP1 was inactivated by genome editing (KEAP1Δ; Extended Data Fig. 3b) we detected a stable BRCA1–PALB2–BRCA2 complex in both G1 and S phases (Fig. 1e). KEAP1 is therefore an inhibitor of the BRCA1–PALB2 interaction.

CUL3 also interacts with PALB2 (Extended Data Fig. 3c), and its depletion in 53BP1Δ U2OS cells de-repressed PALB2 ionizing-radiation-induced foci in G1 (Fig. 1d and Extended Data Fig. 3a). Furthermore, in G1-synchronized KEAP1Δ cells, expression of a CUL3-binding-deficient KEAP1 protein that lacks its BTB domain (ΔBTB) failed to suppress the BRCA1–PALB2 interaction, unlike its wild-type counterpart (Extended Data Fig. 3d). These results suggest that KEAP1 recruits CUL3 to PALB2 to suppress its interaction with BRCA1.

Using the LacR/*lacO* system and co-immunoprecipitation assays, we found that a mutant of PALB2 lacking all eight lysine residues in

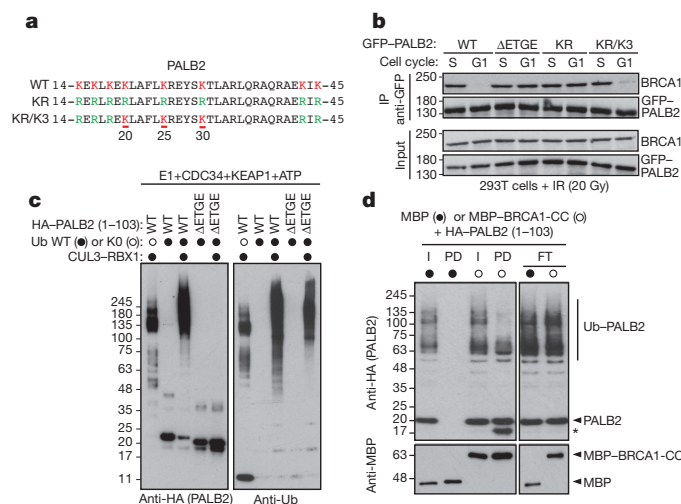
<sup>1</sup>The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, 600 University Avenue, Toronto, Ontario M5G 1X5, Canada. <sup>2</sup>ETH Zurich, Institute of Biochemistry, Department of Biology, Otto-Stern-Weg 3, CH-8093 Zurich, Switzerland. <sup>3</sup>Department of Molecular Genetics, University of Toronto, Ontario M5S 3E1, Canada. <sup>4</sup>Departments of Pathology and Biochemistry & Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada. <sup>5</sup>Department of Radiation Oncology, Rutgers Cancer Institute of New Jersey and Robert Wood Johnson Medical School, Rutgers, The State University of New Jersey, New Brunswick, New Jersey 08901, USA.

\*These authors contributed equally to this work.



**Figure 1 | Inhibition of the BRCA1–PALB2 interaction in G1 is CRL3–KEAP1-dependent.** **a**, Micrographs of irradiated (2 Gy) G1-synchronized U2OS cells processed for  $\gamma$ -H2AX, BRCA1 and BRCA2 immunofluorescence. DAPI, 4',6-diamidino-2-phenylindole; IR, ionizing radiation; WT, wild type. **b**, Quantitation of the experiment shown in **a** and Extended Data Fig. 1d. ASN, asynchronously dividing. Mean  $\pm$  standard deviation (s.d.),  $N = 3$ . **c**, Immunoprecipitation (IP) of PALB2 from extracts prepared from mock- or X-irradiated 293T cells synchronized in S or G1 phases. A normal immunoglobulin (Ig)G

immunoprecipitation was performed as control. Cyclin A staining ascertains cell cycle synchronization. Numbers on left indicate kDa. For gel source data, see Supplementary Fig. 1. **d**, Quantitation of the experiment shown in Extended Data Fig. 3a. **53BP1** $\Delta$  U2OS cells transfected with the indicated GFP–PALB2 vectors and short interfering (si)RNAs were irradiated (20 Gy) before being processed for microscopy (mean  $\pm$  s.d.,  $N = 3$ ). **e**, Normal IgG and PALB2 immunoprecipitations from extracts prepared from synchronized and irradiated 293T cells of the indicated genotypes. Numbers on left indicate kDa.



**Figure 2 | Ubiquitylation of PALB2 prevents BRCA1–PALB2 interaction.** **a**, Sequence of the PALB2 N terminus and mutants. **b**, GFP immunoprecipitation (IP) of extracts derived from G1- or S-phase-synchronized 293T cells expressing the indicated GFP–PALB2 proteins. **c**, *In vitro* ubiquitylation of the indicated HA-tagged PALB2 proteins by CRL3–KEAP1. **d**, Pull-down assay of ubiquitylated HA–PALB2 (1–103) incubated with MBP or MBP–BRCA1–CC. I, input; FT, flow-through; PD, pull-down. The asterisk denotes a fragment of HA–PALB2 competent for BRCA1 binding. **b–d**, Numbers on left indicate kDa.

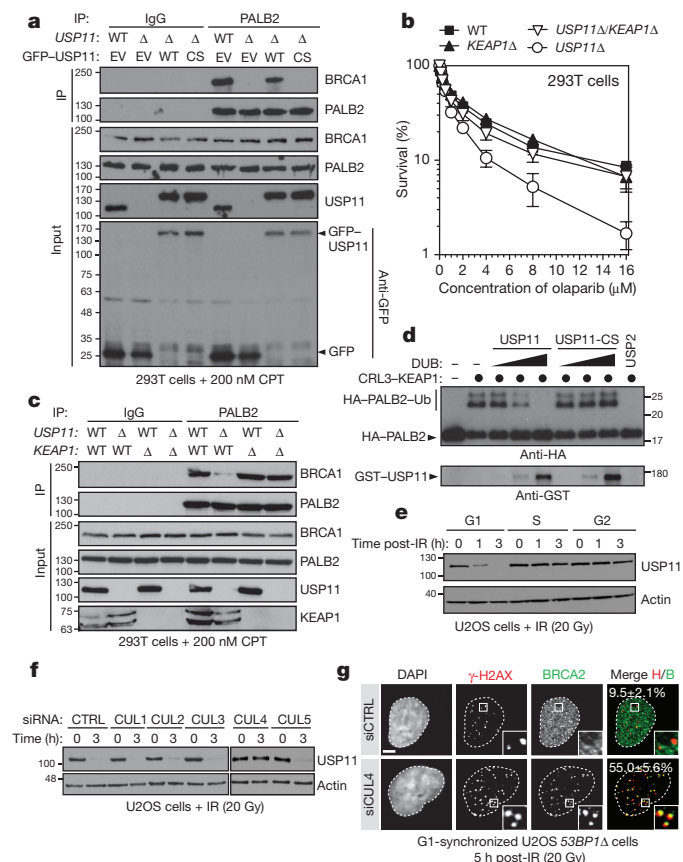
the BRCA1-interaction domain (PALB2-KR; Fig. 2a) could interact with BRCA1 irrespective of cell cycle position (Fig. 2b and Extended Data Fig. 3e, f). Further mutagenesis identified residues 20, 25 and 30 in PALB2 as critical for the suppression of the BRCA1–PALB2 interaction, since reintroduction of these lysines in the context of PALB2–KR (yielding PALB2–KR/K3; Fig. 2a) led to the suppression of BRCA1–PALB2–BRCA2 complex assembly in G1 cells (Fig. 2b and Extended

Data Fig. 3e). Together, these results suggested a model whereby PALB2-bound KEAP1 forms an active CRL3 complex that ubiquitylates the PALB2 N terminus to suppress its interaction with BRCA1.

While PALB2 ubiquitylation can be detected in cells (Extended Data Fig. 4a), the lysine-rich nature of the PALB2 N terminus has so far precluded us from unambiguously mapping *in vivo* ubiquitylation sites on Lys 20, 25 or 30. However, we could detect ubiquitylation on Lys 16 and Lys 43 by mass spectrometry, indicating that the PALB2 N terminus is ubiquitylated (Extended Data Fig. 4b). In a complementary set of experiments, PALB2 targeted to the *lacO* array induced immunoreactivity to conjugated Ub (Extended Data Fig. 4c–e). Ub colocalization with PALB2 was highest in G1, and depended on the KEAP1-interaction motif and the presence of the Lys 20/25/30 residues (Extended Data Fig. 4d–e), consistent with the model that PALB2 is ubiquitylated on those sites in G1 cells. Indeed, we could readily reconstitute ubiquitylation of the N terminus of PALB2 (residues 1–103; fused to a haemagglutinin (HA) epitope tag), by recombinant CRL3–KEAP1, in a manner that depended on the KEAP1-interaction domain of PALB2 (Fig. 2c), and we unambiguously identified Lys 25 and Lys 30 as being ubiquitylated by KEAP1 *in vitro* by mass spectrometry (Extended Data Fig. 5).

Ubiquitylation of PALB2 by CRL3–KEAP1 inhibited its interaction with a BRCA1 fragment comprising residues 1363–1437 (BRCA1-CC), an inhibition that was more obvious with the highly modified forms of PALB2 owing to the presence of ubiquitylated lysines outside the BRCA1-interaction domain (Fig. 2d). To test specifically whether ubiquitylation of a single lysine residue (of the three identified as critical) inhibited the interaction with BRCA1, we used chemical crosslinking to install a single Ub moiety at position 20 or 45 (yielding PALB2–K<sub>20</sub>-Ub and PALB2–K<sub>45</sub>-Ub, respectively). Ubiquitylation of PALB2 at position 20 completely suppressed its interaction with BRCA1 whereas modification of residue 45 had no impact on the interaction (Extended Data Fig. 6a), echoing the *in vivo* data (Extended Data Fig. 3e). Together, these results indicate that ubiquitylation of PALB2 at specific sites on its N terminus prevents its interaction with BRCA1.





**Figure 3 | USP11 opposes the activity of CRL3-KEAP1.** **a**, Normal IgG or PALB2 immunoprecipitation (IP) of extracts derived from camptothecin (CPT)-treated 293T cells of the indicated genotypes transfected with GFP-USP11 constructs. EV, empty vector; CS, C318S; WT, wild type. **b**, Clonogenic survival assays of 293T cells of the indicated genotypes treated with olaparib (mean  $\pm$  s.d.,  $N \geq 3$ ). **c**, Normal IgG or PALB2 immunoprecipitation of extracts derived from CPT-treated 293T cells of the indicated genotypes. **d**, Immunoblots of deubiquitylation reactions containing ubiquitylated HA-tagged PALB2 (1–103) and increasing concentrations of glutathione S-transferase (GST)-USP11 or its C270S (CS) mutant. USP2 was used as a control. DUB, deubiquitylase. **e**, Cell-cycle-synchronized U2OS cells were irradiated (20 Gy dose) and processed for immunoblotting. IR, ionizing radiation. **f**, Immunoblots of extracts from irradiated U2OS cells transfected with the indicated siRNAs. CTRL, control. **g**, Fluorescence micrographs of G1-synchronized and irradiated (20 Gy) 53BP1Δ U2OS cells transfected with the indicated siRNAs. The percentage of cells with more than five  $\gamma$ -H2AX-colocalizing BRCA2 foci is indicated (mean  $\pm$  s.d.,  $N = 3$ ). Scale bars, 5  $\mu$ m. **a**, **c**, **d**, **f**, Numbers to left or right indicate kDa.

Since neither the activity of the CRL3-KEAP1 E3 ligase (Extended Data Fig. 6b) nor the interaction of CRL3-KEAP1 with PALB2 (Extended Data Fig. 3c) are regulated by the cell cycle, we considered the possibility that deubiquitylation of PALB2 might be regulated in a cell-cycle-dependent manner. KEAP1 physically interacts with USP11 (ref. 22), a deubiquitylase that also interacts with BRCA2 (ref. 23) and PALB2 (Extended Data Fig. 6c). USP11 depletion impairs gene conversion<sup>24</sup> (Extended Data Fig. 6d) and results in hypersensitivity to PARP inhibition<sup>24</sup>, identifying it as a homologous recombination regulator of unknown function. Co-immunoprecipitation experiments confirmed that USP11 and its catalytic activity were necessary for the formation of a stable BRCA1-PALB2-BRCA2 complex, especially in the presence of DNA damage (Fig. 3a and Extended Data Fig. 6e, f).

If USP11 antagonizes PALB2 ubiquitylation by CRL3-KEAP1, then removal of KEAP1 (or CUL3) should reverse the phenotypes imparted by loss of USP11. Indeed, deletion of KEAP1 restored resistance to PARP inhibitors (PARPi) and the BRCA1-PALB2 interaction

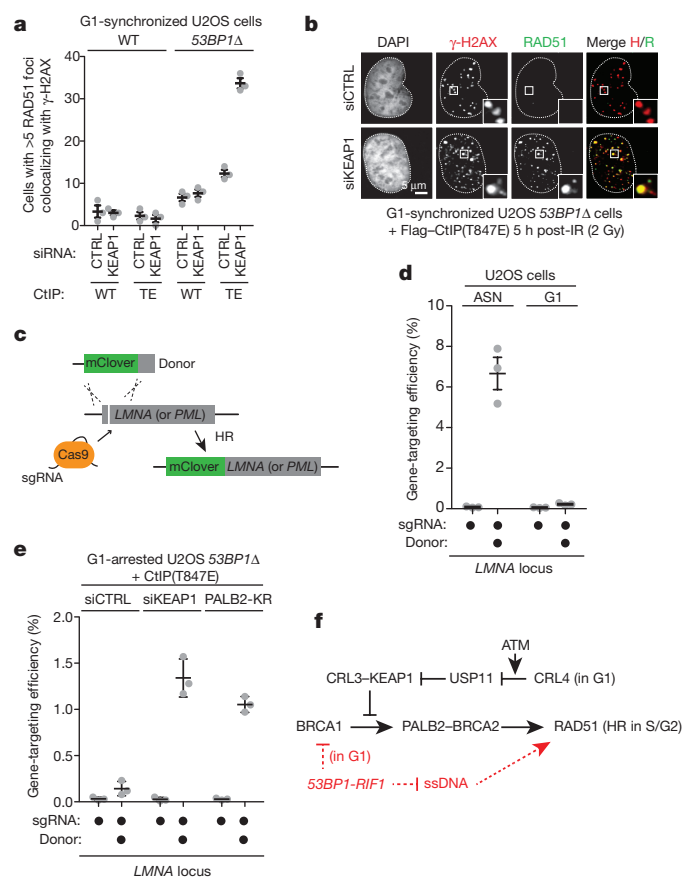
in USP11-knockout cells prepared by genome editing (USP11Δ; Fig. 3b, c and Extended Data Fig. 6e). Likewise, depletion of CUL3 or KEAP1 reversed the gene conversion defect of USP11-depleted cells (Extended Data Fig. 7a). Introduction of the PALB2-KR mutant restored its interaction with BRCA1 and reversed PARPi sensitivity in USP11Δ cells in a manner that depended on Lys 20/25/30 (Extended Data Fig. 7b, c). Since recombinant USP11 can de-ubiquitylate PALB2 (1–103) *in vitro* (Fig. 3d), these results suggest that USP11 promotes the assembly of the BRCA1-PALB2-BRCA2 complex by reversing the inhibitory ubiquitylation on the PALB2 Lys 20/25/30 residues.

We observed that USP11 turns over rapidly in G1 cells and interacts poorly with PALB2 in this phase of the cell cycle (Extended Data Fig. 8a, b). Furthermore, there is a rapid loss of USP11 upon DNA damage induction, specifically in G1 phase (Fig. 3e and Extended Data Fig. 8b, c). The destabilization of USP11 after ionizing radiation treatment is dependent on ATM signalling, whereas it is ATR-dependent after ultraviolet irradiation (Extended Data Fig. 8d, e). The drop in USP11 steady-state levels in G1 is the result of proteasomal degradation (Extended Data Fig. 8f). A CRL4 E3 Ub ligase is probably responsible for controlling the stability of USP11, as treatment with MLN4924, a pan-CRL inhibitor<sup>25</sup> (Extended Data Fig. 8g), or depletion of CUL4 (Fig. 3f), protected USP11 from DNA-damage-induced degradation. CUL4 depletion led to BRCA2 and PALB2 ionizing-radiation-induced focus formation in G1 53BP1Δ cells (Fig. 3g and Extended Data Fig. 9a), consistent with the regulation of USP11 by a CRL4 complex acting as the upstream signal that ultimately controls BRCA1-PALB2-BRCA2 complex assembly.

While deletion of 53BP1 produces low levels of ssDNA in G1 cells<sup>26</sup>, combining the 53BP1Δ mutation with depletion of KEAP1 did not produce extraction-resistant RAD51 ionizing-radiation-induced foci, suggesting little-to-no RAD51 nucleofilament formation (Extended Data Fig. 9b). We surmised that ssDNA formation remained insufficient in those cells and thus took advantage of the phosphomimetic T847E mutant of CtIP that promotes resection in G1 cells<sup>27</sup>. Unlike wild-type CtIP, introduction of CtIP(T847E) into 53BP1Δ cells depleted of KEAP1 induced RAD51 ionizing-radiation-induced focus formation in G1 cells (Fig. 4a, b and Extended Data Fig. 9b, c) along with unscheduled DNA synthesis (Extended Data Fig. 9d). These results suggested that the steps downstream of RAD51 nucleofilament formation, that is, strand invasion, D-loop formation and DNA synthesis, could be activated in G1.

To test whether productive homologous recombination could also be activated in G1, we employed a CRISPR-Cas9-stimulated gene-targeting assay<sup>28</sup> in which the insertion of the coding sequence for the mClover fluorescent protein at the 5' end of the lamin A (*LMNA*) or *PML* genes was monitored by microscopy or flow cytometry (Fig. 4c and Extended Data Fig. 9e, f), with the latter method enabling the gating of cells with a defined DNA content (such as G1 cells). We also established synchronization protocols in which G1 cells obtained after release from a thymidine block were arrested in G1 by lovastatin treatment<sup>2</sup> for 24 h (Extended Data Fig. 9g, h). Using this system, we determined a concentration of donor template in the linear range of the assay and ascertained that gene targeting at the *LMNA* locus was dependent on BRCA1-PALB2-BRCA2 complex assembly (Extended Data Fig. 10a, b). We also confirmed that gene targeting by homologous recombination was highly suppressed in G1 (Fig. 4d).

The combined activation of resection and BRCA1 recruitment to DSB sites (that is, in 53BP1Δ cells expressing CtIP(T847E)) was insufficient to stimulate gene targeting at either the *LMNA* or the *PML* locus in G1 cells (Fig. 4e and Extended Data Fig. 10c, d). However, when the BRCA1-PALB2 interaction was restored in resection-competent G1 cells using either KEAP1 depletion or expression of the PALB2-KR mutant, we detected a robust increase in gene-targeting events at both loci (Fig. 4e and Extended Data Fig. 10c, d). We note, however, that the gene-targeting frequencies of G1 cells remained lower than those of asynchronously dividing cells, suggesting an incomplete activation of



**Figure 4 | Reactivation of homologous recombination in G1.**

**a**, Quantitation of wild-type (WT) and *53BP1*Δ U2OS cells co-transfected with non-targeting (CTRL) or KEAP1 siRNAs and vectors expressing wild-type CtIP or the T847E (TE) mutant that were synchronized in G1, irradiated (2 Gy) and processed for γ-H2AX and RAD51 immunofluorescence (mean ± s.d., *N* = 3). **b**, Representative micrographs from **a**. IR, ionizing radiation. **c**, Schematic of the gene-targeting assay. **d**, Gene-targeting efficiency at the *LMNA* locus in asynchronously dividing (ASN) and G1-arrested U2OS cells (mean ± s.d., *N* = 3). HR, homologous recombination; sgRNA, single guide RNA. **e**, Gene targeting at the *LMNA* locus in G1-arrested cells transfected with the indicated siRNA or a PALB2-KR expression vector (mean ± s.d., *N* = 3). **f**, Model of the cell-cycle regulation of homologous recombination.

homologous recombination. 53BP1 inactivation and the expression of CtIP(T847E) were both necessary for G1 homologous recombination (Extended Data Fig. 10e, f), indicating that the simultaneous activation of end resection and BRCA2 recruitment to DSB sites were both necessary and sufficient to activate unscheduled recombination in this phase of the cell cycle.

We conclude that the regulation of BRCA1–PALB2–BRCA2 complex assembly is a key node in the cell cycle control of DSB repair by homologous recombination. This regulation converges on the BRCA1-interaction site on PALB2 and is enforced by the opposing activities of the E3 ligase CRL3–KEAP1 and the deubiquitylase USP11, with the latter being antagonized in G1 by a CRL4 complex (Fig. 4f). Our studies also demonstrate that the suppression of homologous recombination in G1 cells is largely reversible and that it involves the combined suppression of end resection and BRCA2 recruitment to DSB sites (Fig. 4f). As most cells in the human body are not actively cycling and are thus refractory to homologous recombination, the manipulations described here may eventually enable therapeutic gene targeting in a wide variety of tissues. However, these approaches may necessitate the reversal of additional blocks to gene targeting such as the potential downregulation of homologous recombination factor expression in post-mitotic cells.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 7 April; accepted 13 October 2015.**

**Published online 9 December 2015.**

- Jasin, M. & Rothstein, R. Repair of strand breaks by homologous recombination. *Cold Spring Harb. Perspect. Biol.* **5**, a012740 (2013).
- Hartlerode, A., Odate, S., Shim, I., Brown, J. & Scully, R. Cell cycle-dependent induction of homologous recombination by a tightly regulated I-SceI fusion protein. *PLoS ONE* **6**, e16501 (2011).
- Rothkamm, K., Krüger, I., Thompson, L. H. & Löbrich, M. Pathways of DNA double-strand break repair during the mammalian cell cycle. *Mol. Cell. Biol.* **23**, 5706–5715 (2003).
- Panier, S. & Durocher, D. Push back to respond better: regulatory inhibition of the DNA double-strand break response. *Nature Rev. Mol. Cell Biol.* **14**, 661–672 (2013).
- Ma, J. et al. PALB2 interacts with KEAP1 to promote NRF2 nuclear accumulation and function. *Mol. Cell. Biol.* **32**, 1506–1517 (2012).
- Genschik, P., Sumara, I. & Lechner, E. The emerging family of CULLIN3-RING ubiquitin ligases (CRL3s): cellular functions and disease implications. *EMBO J.* **32**, 2307–2320 (2013).
- Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nature Rev. Cancer* **12**, 68–78 (2012).
- Li, M. L. & Greenberg, R. A. Links between genome integrity and BRCA1 tumor suppression. *Trends Biochem. Sci.* **37**, 418–424 (2012).
- Park, J. Y., Zhang, F. & Andreassen, P. R. PALB2: the hub of a network of tumor suppressors involved in DNA damage responses. *Biochim. Biophys. Acta* **1846**, 263–275 (2014).
- Zhang, F. et al. PALB2 links BRCA1 and BRCA2 in the DNA-damage response. *Curr. Biol.* **19**, 524–529 (2009).
- Sy, S. M., Huen, M. S. & Chen, J. PALB2 is an integral component of the BRCA complex required for homologous recombination repair. *Proc. Natl Acad. Sci. USA* **106**, 7155–7160 (2009).
- Simhadri, S. et al. Male fertility defect associated with disrupted BRCA1–PALB2 interaction in mice. *J. Biol. Chem.* **289**, 24617–24629 (2014).
- Bhattacharyya, A., Ear, U. S., Koller, B. H., Weichselbaum, R. R. & Bishop, D. K. The breast cancer susceptibility gene *BRCA1* is required for subnuclear assembly of Rad51 and survival following treatment with the DNA cross-linking agent cisplatin. *J. Biol. Chem.* **275**, 23899–23903 (2000).
- Zhang, F., Bick, G., Park, J. Y. & Andreassen, P. R. MDC1 and RNF8 function in a pathway that directs BRCA1-dependent localization of PALB2 required for homologous recombination. *J. Cell Sci.* **125**, 6049–6057 (2012).
- Escribano-Diaz, C. et al. A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. *Mol. Cell* **49**, 872–883 (2013).
- Feng, L., Fong, K. W., Wang, J., Wang, W. & Chen, J. RIF1 counteracts BRCA1-mediated end resection during DNA repair. *J. Biol. Chem.* **288**, 11135–11143 (2013).
- Chapman, J. R. et al. RIF1 is essential for 53BP1-dependent nonhomologous end joining and suppression of DNA double-strand break resection. *Mol. Cell* **49**, 858–871 (2013).
- Bunting, S. F. et al. 53BP1 inhibits homologous recombination in *Brca1*-deficient cells by blocking resection of DNA breaks. *Cell* **141**, 243–254 (2010).
- Zimmermann, M., Lottersberger, F., Buonomo, S. B., Sfeir, A. & de Lange, T. 53BP1 regulates DSB repair using Rif1 to control 5' end resection. *Science* **339**, 700–704 (2013).
- Tang, J. et al. Acetylation limits 53BP1 association with damaged chromatin to promote homologous recombination. *Nature Struct. Mol. Biol.* **20**, 317–325 (2013).
- Taguchi, K., Motohashi, H. & Yamamoto, M. Molecular mechanisms of the Keap1–Nrf2 pathway in stress response and cancer evolution. *Genes Cells* **16**, 123–140 (2011).
- Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389–403 (2009).
- Schoenfeld, A. R., Apgar, S., Dolios, G., Wang, R. & Aaronson, S. A. BRCA2 is ubiquitinated *in vivo* and interacts with USP11, a deubiquitinating enzyme that exhibits prosurvival function in the cellular response to DNA damage. *Mol. Cell. Biol.* **24**, 7444–7455 (2004).
- Wiltshire, T. D. et al. Sensitivity to poly(ADP-ribose) polymerase (PARP) identifies ubiquitin-specific peptidase 11 (USP11) as a regulator of DNA double-strand break repair. *J. Biol. Chem.* **285**, 14565–14571 (2010).
- Enchev, R. I., Schulman, B. A. & Peter, M. Protein neddylation: beyond cullin-RING ligases. *Nature Rev. Mol. Cell Biol.* **16**, 30–44 (2015).
- Yamane, A. et al. RPA accumulation during class switch recombination represents 5'–3' DNA-end resection during the S-G2/M phase of the cell cycle. *Cell Reports* **3**, 138–147 (2013).
- Huertas, P. & Jackson, S. P. Human CtIP mediates cell cycle control of DNA end resection and double strand break repair. *J. Biol. Chem.* **284**, 9558–9565 (2009).
- Pinder, J., Salsman, J. & Delliare, G. Nuclear domain 'knock-in' screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic Acids Res.* **43**, 9379–9392 (2015).



**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We are grateful to R. Szilard and X.-D. Zhu for critical reading of the manuscript; to D. Lo, M. Canny and J. Young for help on the project. We also thank B. Larsen and M. Tucholska for technical support, J. Stark for the U2OS DR-GFP cells, R. Greenberg for the U2OS 256 cells, F. Sicheri for ubiquitin reagents, F. Shao for the *KEAP1* bacterial expression vector and D. Cortez for *USP11* cDNA. A.O. is a Scholar of the Terry Fox Foundation Strategic Training Initiative for Excellence in Radiation Research for the 21st Century (EIRR21); S.M.N. receives a postdoctoral fellowship from the Dutch Cancer Society (KWF); M.D.W. holds a long-term Human Frontier Science Program fellowship; A.S. receives an Ontario Graduate Scholarship. R.I.E. was funded by a Marie Curie postdoctoral fellowship. J.P. was supported by the Beatrice Hunter Cancer Research Institute (BHCRI) with funds provided by the Harvey Graham Cancer Research Fund as part of the Terry Fox Foundation Strategic Health Research Training Program in Cancer Research at the Canadian Institutes of Health Research (CIHR). G.D. is a Senior Scientist of the BHCRI. D.D. is the Thomas Kierans Chair in Mechanisms of Cancer Development and a Canada

Research Chair (Tier 1) in the Molecular Mechanisms of Genome Integrity. Work was supported by a Grant-in-Aid from the Krembil Foundation (to D.D.) and CIHR grants FDN143343 (to D.D.) and MOP84260 (to G.D.).

**Author Contributions** A.O. carried out all cellular experiments. S.M.N. carried out *in vitro* ubiquitin-related experiments and mass spectrometry. M.D.W. produced recombinant KEAP1, USP11 and chemically ubiquitylated PALB2. S.L. produced the *53BP1*Δ cells. R.I.E. produced neddylated CUL3–RBX1. A.S. and M.M. helped A.O. B.X. contributed PALB2 reagents and advice. J.P., J.S. and G.D. provided reagents and advice for the gene-targeting assay. M.P. supervised R.I.E. D.D. supervised the project and wrote the manuscript with A.O. and S.M.N., with input from the other authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.D. ([durocher@lunenfeld.ca](mailto:durocher@lunenfeld.ca)).

## METHODS

**Plasmids.** The complementary DNA of *PALB2* was obtained from the Mammalian Gene Collection (MGC). Full-length *PALB2* and *BRCA1* were amplified by PCR, subcloned into pDONR221 and delivered into the pDEST-GFP, pDEST-Flag and the mCherry-LacR vectors using Gateway cloning technology (Invitrogen). Similarly, the coiled-coil domain of *BRCA1* (residues 1363–1437) was amplified by PCR, subcloned into the pDONR221 vector and delivered into both mCherry-LacR and pDEST-GFP vectors. The N-terminal domain of *PALB2* was amplified by PCR and introduced into the GST expression vector pET30-2-His-GST-TEV<sup>29</sup> using the EcoRI/XhoI sites. The coiled-coil domain of *BRCA1* was cloned into pMAL-c2 using the BamHI/SalI sites. Truncated forms of *PALB2* were obtained by introducing stop codons or deletions through site-directed mutagenesis. Full-length CtIP was amplified by PCR, subcloned into the pDONR221 and delivered into the lentiviral construct pCW57.1 (a gift from D. Root; Addgene plasmid #41393) using Gateway cloning technology (Invitrogen). The *USP11* cDNA was a gift from D. Cortez and was amplified by PCR and cloned into the pDsRed2-C1 vector using the EcoRI/SalI sites. The bacterial codon-optimized coding sequence of pig *USP11* was subcloned into the 6×His-GST vector pETM-30-Htb using the BamHI/EcoRI sites. siRNA-resistant versions of *PALB2*, *BRCA1* and *USP11* constructs were generated as previously described<sup>11</sup>. Full-length *CUL3* and *RBX1* were amplified by PCR from a human pancreas cDNA library (Invitrogen) as previously described<sup>30</sup> and cloned into the dual expression pFBDM vector using NheI/XmaI and BssHII/NotI respectively. The *NEDD8* cDNA was a gift from D. Xirodimas and was fused to a double StrepII tag at its C terminus in the pET17b vector (Millipore). Human *DEN1* was amplified from a vector supplied by A. Echalié and fused to a non-cleavable N-terminal StrepII2× tag by PCR and inserted into a pET17b vector. The pCOOL-mKEAP1 plasmid was a gift from F. Shao. The pcDNA3-HA2-KEAP1 and pcDNA3-HA2-KEAP1ΔBTB were gifts from Y. Xiong (Addgene plasmids #21556 and 21593). gRNAs were synthesized and processed as described previously<sup>31</sup>. Annealed gRNAs were cloned into the Cas9-expressing vectors pSpCas9(BB)-2A-Puro (PX459) or pX330-U6-Chimeric\_BB-CBH-hSpCas9, a gift from F. Zhang (Addgene plasmids #48139 and 42230). The gRNAs targeting the *LMNA* or the *PML* locus and the mClover-tagged *LMNA* or *PML* are described previously<sup>28</sup>. The lentiviral packaging vector psPAX2 and the envelope vector VSV-G were a gift from D. Trono (Addgene plasmids #12260 and 12259). His<sub>6</sub>-Ub was cloned into the pcDNA5-FRT/TO backbone using the XhoI/HindIII sites. All mutations were introduced by site-directed mutagenesis using QuickChange (Stratagene) and all plasmids were sequence-verified.

**Cell culture and plasmid transfection.** All culture media were supplemented with 10% fetal bovine serum (FBS). U-2-OS (U2OS) cells were cultured in McCoy's medium (Gibco). 293T cells were cultured in DMEM (Gibco). Parental cells were tested for mycoplasma contamination and authenticated by STR DNA profiling. Plasmid transfections were carried out using Lipofectamine 2000 Transfection Reagent (Invitrogen) following the manufacturer's protocol. Lentiviral infection was carried out as previously described<sup>15</sup>. U2OS and 293T cells were purchased from ATCC. U2OS 256 cells were a gift from R. Greenberg.

**Antibodies.** We employed the following antibodies: rabbit anti-53BP1 (A300-273A, Bethyl), rabbit anti-53BP1 (sc-22760, Santa Cruz), mouse anti-53BP1 (#612523, BD Biosciences), mouse anti-γ-H2AX (clone JBW301, Millipore), rabbit anti-γ-H2AX (#2577, Cell Signaling Technologies), rabbit anti-KEAP1 (ab66620, Abcam), rabbit anti-NRF2 (ab62352, Abcam), mouse anti-Flag (clone M2, Sigma), mouse anti-tubulin (CP06, Calbiochem), mouse anti-GFP (#11814460001, Roche), mouse anti-CCNA (MONX10262, Monosan), rabbit anti-BRCA2 (ab9143, Abcam), mouse anti-BRCA2 (OP95, Calbiochem), rabbit anti-BRCA1 (#07-434, Millipore), rabbit anti-USP11 (ab109232, Abcam), rabbit anti-USP11 (A301-613A, Bethyl), rabbit anti-RAD51 (#70-001, Bioacademia), mouse anti-BrdU (RPN202, GE Healthcare), mouse anti-FK2 (BML-PW8810, Enzo), rabbit anti-PALB2 (ref. 32), rabbit anti-GST (sc-459, Santa Cruz), rabbit anti-CUL3 (A301-108A, Bethyl), mouse anti-MBP (E8032, NEB), mouse anti-HA (clone 12CA5, a gift from M. Tyers), rabbit anti-ubiquitin (Z0458, Dako) and mouse anti-actin (CP01, Calbiochem). The following antibodies were used as secondary antibodies in immunofluorescence microscopy: Alexa Fluor 488 donkey anti-rabbit IgG, Alexa Fluor 488 donkey anti-goat IgG, Alexa Fluor 555 donkey anti-mouse IgG, Alexa Fluor 555 donkey anti-rabbit IgG, Alexa Fluor 647 donkey anti-mouse IgG, Alexa Fluor 647 donkey anti-human IgG, Alexa Fluor 647 donkey anti-goat IgG (Molecular Probes).

**RNA interference.** All siRNAs employed in this study were single duplex siRNAs purchased from ThermoFisher. RNA interference (RNAi) transfections were performed using Lipofectamine RNAiMax (Invitrogen) in a forward transfection mode. The individual siRNA duplexes used were: *BRCA1* (D-003461-05), *PALB2* (D-012928-04), *USP11* (D-006063-01), *CUL1* (M-004086-01), *CUL2* (M-007277-00), *CUL3* (M-0010224-02), *CUL4A* (M-012610-01), *CUL4B* (M-017965-01), *CUL5*

(M-019553-01), *KEAP1* (D-12453-02), *RAD51* (M-003530-04), *ChIP/RBBP8* (M-001376-00), *BRCA2* (D-003462-04), *53BP1* (D-003549-01) and non-targeting control siRNA (D-001210-02). Except when stated otherwise, siRNAs were transfected 48 h before cell processing.

**Inhibitors and fine chemicals.** We employed the following drugs at the indicated concentrations: cycloheximide (CHX; Sigma) at 100 ng ml<sup>-1</sup>, camptothecin (CPT; Sigma) at 0.2 μM, ATM inhibitor (KU55933; Selleck Chemicals) at 10 μM, ATR inhibitor (VE-821; a gift from P. Reaper) at 10 μM, DNA-PKcs inhibitor (NU7441; Genetex) at 10 μM, proteasome inhibitor MG132 (Sigma) at 2 μM, lovastatin (S2061; Selleck Chemicals) at 40 μM, doxycycline (#8634-1; Clontech), Nedd8-activating enzyme inhibitor (MLN4929; Active Biochem) at 5 μM and olaparib (Selleck) at the indicated concentrations.

**Immunofluorescence microscopy.** In most cases, cells were grown on glass coverslips, fixed with 2% (w/v) paraformaldehyde in PBS for 20 min at room temperature, permeabilized with 0.3% (v/v) Triton X-100 for 20 min at room temperature and blocked with 5% BSA in PBS for 30 min at room temperature. Alternatively, cells were fixed with 100% cold methanol for 10 min at -20°C and subsequently washed with PBS for 5 min at room temperature before PBS-BSA blocking. Cells were then incubated with the primary antibody diluted in PBS-BSA for 2 h at room temperature. Cells were next washed with PBS and then incubated with secondary antibodies diluted in PBS-BSA supplemented with 0.8 μg ml<sup>-1</sup> of DAPI to stain DNA for 1 h at room temperature. The coverslips were mounted onto glass slides with Prolong Gold mounting agent (Invitrogen). Confocal images were taken using a Zeiss LSM780 laser-scanning microscope. For G1 versus S/G2 analysis of the *BRCA1*-*PALB2*-*BRCA2* axis, cells were first synchronized with a double-thymidine block, released to allow entry into S phase and exposed to 2 or 20 Gy of X-irradiation at 5 h and 12 h post-release and fixed at 1 to 5 h post-treatment (where indicated). For the examination of DNA replication, cells were pre-incubated with 30 μM BrdU for 30 min before irradiation and processed as previously described.

**CRISPR-Cas9 genome editing of USP11/KEAP1.** 293T and U2OS cells were transiently transfected with three distinct sgRNAs targeting either 53BP1, USP11 or KEAP1 and expressed from the pX459 vector containing Cas9 followed by the 2A-Puromycin cassette. The next day, cells were selected with puromycin for 2 days and subcloned to form single colonies or subpopulations. Clones were screened by immunoblot and/or immunofluorescence to verify the loss of 53BP1, USP11 or KEAP1 expression and subsequently characterized by PCR and sequencing. The genomic region targeted by the CRISPR-Cas9 was amplified by PCR using Turbo Pfu polymerase (Agilent) and the PCR product was cloned into the pCR2.1 TOPO vector (Invitrogen) before sequencing.

**Olaparib clonogenic assay.** 293T cells were incubated with the indicated doses of olaparib (Selleck Chemicals) for 24 h, washed once with PBS and counted by trypan blue staining. Five-hundred cells were then plated in duplicate for each condition. The cell survival assay was performed as previously described<sup>33</sup>.

**Recombinant protein production.** GST and MBP fusions proteins were produced as previously described<sup>34,35</sup>. Briefly, MBP proteins expressed in *Escherichia coli* were purified on amylose resin (New England Biolabs) according to the batch method described by the manufacturer and stored in 1× PBS, 5% glycerol. GST proteins expressed in *E. coli* were purified on glutathione sepharose 4B (GE Healthcare) resin in 50 mM Tris HCl pH 7.5, 300 mM NaCl, 2 mM dithiothreitol (DTT), 1 mM EDTA, 15 μg ml<sup>-1</sup> AEBSE and 1× complete protease inhibitor cocktail (Roche). Upon elution from the resin using 50 mM glutathione in 50 mM Tris HCl pH 8, 2 mM DTT, the His<sub>6</sub>-GST tag was cleaved off using His-tagged TEV protease (provided by F. Sicheri) in 50 mM Tris HCl pH 7.5, 150 mM NaCl, 10 mM glutathione, 10% glycerol, 2 mM sodium citrate and 2 mM β-mercaptoethanol. His<sub>6</sub>-tagged proteins were depleted using Ni-NTA-agarose beads (Qiagen) in 50 mM Tris HCl pH 7.5, 300 mM NaCl, 20 mM imidazole, 5 mM glutathione, 10% glycerol, 1 mM sodium citrate and 2 mM β-mercaptoethanol followed by centrifugal concentration (Amicon centrifugal filters, Millipore). GST-mKEAP1 was purified as described previously<sup>36</sup>, with an additional anion exchange step on a HiTrap Q HP column (GE Healthcare). The GST tag was left on the protein for *in vitro* experiments. Purification of *CUL3* and *RBX1* was performed as previously described<sup>30</sup>. NEDD8 (gift from D. Xirodimas) and *DEN1* were expressed in *E. coli* BL21 grown in Terrific broth media and induced overnight with 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) at 16°C. Cells were harvested and resuspended in wash buffer (400 mM NaCl, 50 mM Tris-HCl, pH 8, 5% glycerol, 2 mM DTT), supplemented with lysozyme, universal nuclease (Pierce), benzamide, leupeptin, pepstatin, PMSF and complete protease inhibitor cocktail (Roche), except for *DEN1*-expressing cells where the protease inhibitors were omitted. Cells were lysed by sonication and the lysate was cleared by centrifugation at 20,000 r.p.m. for 50 min. The soluble supernatant was bound to a 5 ml Strep-Tactin Superflow Cartridge with a flow rate of 3 ml min<sup>-1</sup> using a peristaltic pump.

The column was washed with 20 column volumes (CV) of washing buffer and eluted with 5 CV washing buffer, diluted 1:2 in water to reduce the final salt concentration, and supplemented with 2.5 mM desthiobiotin. The elution fractions were pooled and concentrated to a total volume of 4 ml using a 3 kDa cut-off Amicon concentrator. DEN1 was further purified over a Superdex 75 size-exclusion column, buffer exchanged into 150 mM NaCl, HEPES, pH 7.6, 2% glycerol and 1 mM DTT. The C-terminal pro-peptide and StrepII2 $\times$ -tag were removed by incubation with StrepII2 $\times$ -DEN1 in a 1:20 molar ratio for 1 h at room temperature. The DEN1 cleavage reaction was buffer exchanged on a Zeba MWCO desalting column (Pierce), to remove the desthiobiotin, and passed through a Strep-Tactin Cartridge, which retains the C-terminal pro-peptide and DEN1. The GST-tagged *Sus scrofa* (pig) USP11 proteins were expressed in *E. coli* as described<sup>37</sup>. Cells were lysed by lysozyme treatment and sonication in 50 mM Tris pH 7.5, 300 mM NaCl, 1 mM EDTA, 1 mM AEBSE, 1 $\times$  Protease Inhibitor mix (284 ng ml<sup>-1</sup> leupeptin, 1.37  $\mu$ g ml<sup>-1</sup> pepstatin A, 170  $\mu$ g ml<sup>-1</sup> PMSF and 330  $\mu$ g ml<sup>-1</sup> benzamidin) and 5% glycerol. Cleared lysate was applied to a column packed with glutathione sepharose 4B (GE Healthcare), washed extensively with lysis buffer before elution in 50 mM Tris pH 7.5, 150 mM NaCl, 5% glycerol and 25 mM reduced glutathione. DUB activity was assayed on fluorogenic ubiquitin-AMC (Enzo life sciences), measured using a Synergy Neo microplate reader (Biotek). His<sub>6</sub>-TEV-ubiquitin-G76C was purified on chelating HiTrap resin, following the manufacturer's instructions, followed by size-exclusion chromatography on a S-75 column (GE healthcare). The protein was extensively dialysed in 1 mM acetic acid and lyophilized.

**In vitro ubiquitylation and deubiquitylation of PALB2.** HA-tagged N-terminal fragments of PALB2 (1–103) (1  $\mu$ M) were *in vitro* ubiquitylated using 50  $\mu$ M wild-type (Ubi WT, Boston Biochem) or a lysine-less ubiquitin (Ub-K0, Boston Biochem), 100 nM human UBA1 (E1), 500 nM CDC34 (provided by F. Sicheri and D. Ceccarelli), 250 nM neddylated CUL3/RBX1, 375 nM GST-mKEAP1 and 1.5 mM ATP in a buffer containing 50 mM Tris HCl pH 7.5, 20 mM NaCl, 10 mM MgCl<sub>2</sub> and 0.5 mM DTT. Ubiquitylation reactions were carried out at 37°C for 1 h, unless stated otherwise. For USP11-mediated deubiquitylation assays, HA-PALB2 (1–103) was first ubiquitylated using lysine-less ubiquitin with enzyme concentrations as described earlier in 50  $\mu$ l reactions in a buffer containing 25 mM HEPES pH 8, 150 mM NaCl, 10 mM MgCl<sub>2</sub>, 0.5 mM DTT and 1.5 mM ATP for 1.5 h at 37°C. Reactions were stopped by the addition of 1 unit Apyrase (New England Biolabs). Reaction products were mixed at a 1:1 ratio with wild-type or catalytically inactive (C270S) USP11, or USP2 (provided by F. Sicheri and E. Zeqiraj) using final concentrations of 100 nM, 500 nM and 2,500 nM (USP11) and 500 nM (USP2) and incubated for 2 h at 30°C in a buffer containing 25 mM HEPES pH 8, 150 mM NaCl, 2 mM DTT, 0.1 mg ml<sup>-1</sup> BSA, 0.03% Brij-35, 5 mM MgCl<sub>2</sub>, 0.375 mM ATP. **Pulldown experiments between purified PALB2 and BRCA1.** PALB2 *in vitro* ubiquitylation reaction products were diluted in a buffer at final concentration of 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM MgCl<sub>2</sub>, 0.25 mM DTT and 0.1% NP-40. Twenty micrograms MBP or MBP-BRCA1-CC was coupled to amylose resin (New England Biolabs) in the above buffer supplemented with 0.1% BSA before addition of the ubiquitylation products. Pulldown reactions were performed at 4°C for 2 h, followed by extensive washing.

**Co-immunoprecipitation.** Cells were collected by trypsinization, washed once with PBS and lysed in 500  $\mu$ l of lysis buffer (20 mM Tris-HCl pH 8.0, 150 mM NaCl, 10% glycerol, 2 mM EDTA, 1% NP-40, complete protease inhibitor cocktail (Roche), cocktail of phosphatase inhibitors (Sigma) and N-ethylmaleimide to inhibit deubiquitylation) on ice. Lysates were centrifuged at 15,000g for 10 min at 4°C and protein concentration was evaluated using absorbance at 280 nm. Equivalent amounts of proteins (~0.5–1 mg) were incubated with 2  $\mu$ g of rabbit anti-PALB2, rabbit anti-USP11 antibody, rabbit anti-GFP antibody or normal rabbit IgG for 5 h at 4°C. A mix of protein A/protein G-Sepharose beads (Thermo Scientific) was added for an additional hour. Beads were collected by centrifugation, washed twice with lysis buffer and once with PBS, and eluted by boiling in 2 $\times$  Laemmli buffer before analysis by SDS-PAGE and immunoblotting. For mass spectrometry analysis of Flag-PALB2, 150  $\times$  10<sup>6</sup> transiently transfected HEK293T cells were lysed in high-salt lysis buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100, 3 mM MgCl<sub>2</sub>, 3 mM CaCl<sub>2</sub>), supplemented with complete protease inhibitor cocktail (Roche), 4 mM 1,10-Phenanthroline, 50 U benzonase and 50 U micrococcal nuclease. Cleared lysates were incubated with Flag-M2 agarose (Sigma), followed by extensive washing in lysis buffer and 50 mM ammoniumbicarbonate.

**Mass spectrometry.** After immunoprecipitation of transiently transfected Flag-PALB2 from siCTRL-transfected or USP11 siRNA-depleted 293T cells, cysteine residues were reduced and alkylated on beads using 10 mM DTT (30 min at 56°C) and 15 mM 2-chloroacetamide (1 h at room temperature), respectively. Proteins were digested using limited trypsin digestion on beads (1  $\mu$ g trypsin; Worthington) per sample, 20 min at 37°C, and dried to completeness. For LC-MS/MS analysis,

peptides were reconstituted in 5% formic acid and loaded onto a 12 cm fused silica column with pulled tip packed in-house with 3.5  $\mu$ m Zorbax C18 (Agilent Technologies). Samples were analysed using an Orbitrap Velos (Thermo Scientific) coupled to an Eksigent nanoLC ultra (AB SCIEX). Peptides were eluted from the column using a 90 min linear gradient from 2% to 35% acetonitrile in 0.1% formic acid. Tandem MS spectra were acquired in a data-dependent mode for the top two most abundant multiply charged peptides and included targeted scans for five specific N-terminal PALB2 tryptic digest peptides (charge state 1+, 2+, 3+), either in non-modified form or including a diGly-ubiquitin trypsin digestion remnant. Tandem MS spectra were acquired using collision-induced dissociation. Spectra were searched against the human Refseq\_V53 database using Mascot, allowing up to four missed cleavages and including carbamidomethyl (C), deamidation (NQ), oxidation (M), GlyGly (K) and LeuArgGlyGly (K) as variable modifications.

*In vitro* ubiquitylated HA-PALB2 (1–103) (50  $\mu$ l total reaction mix) was run briefly onto an SDS-PAGE gel, followed by total lane excision, in-gel reduction using 10 mM DTT (30 min at 56°C), alkylation using 50 mM 2-chloroacetamide and trypsin digestion for 16 h at 37°C. Digested peptides were mixed with 20  $\mu$ l of a mix of 10 unique heavy isotope-labelled N-terminal PALB2 (AQUA) peptides (covering full or partial tryptic digests of regions surrounding Lys 16, 25, 30 or 43, either in non-modified or diG-modified form; 80–1,200 fmol  $\mu$ l<sup>-1</sup> per peptide, based on individual peptide sensitivity testing) before loading 6  $\mu$ l onto a 12 cm fused silica column with pulled tip packed in-house with 3.5  $\mu$ m Zorbax C18. Samples were measured on an Orbitrap ELITE (Thermo Scientific) coupled to an Eksigent nanoLC ultra (AB SCIEX). Peptides were eluted from the column using a 180 min linear gradient from 2% to 35% acetonitrile in 0.1% formic acid. Tandem MS spectra were acquired in a data-dependent mode for the top two most abundant multiply charged ions and included targeted scans for the ten specific N-terminal PALB2 tryptic digest peptides (charge states 1+, 2+, 3+), either in light or heavy isotope-labelled form. Tandem MS spectra were acquired using collision induced dissociation. Spectra were searched against the human Refseq\_V53 database using Mascot, allowing up to two missed cleavages and including carbamidomethyl (C), deamidation (NQ), oxidation (M), GlyGly (K) and LeuArgGlyGly (K) as variable modifications, after which spectra were manually validated.

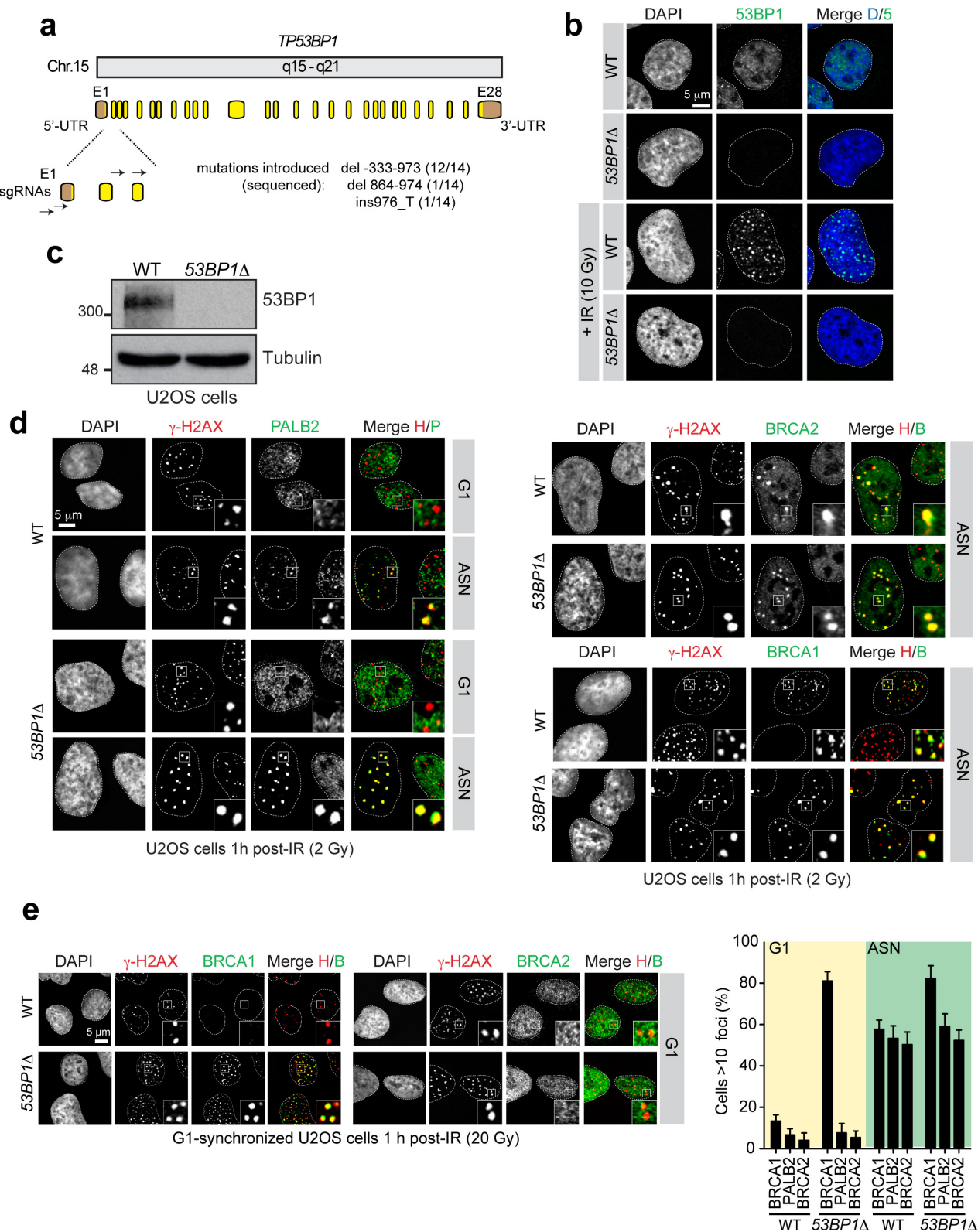
**His-Ub pulldown.** 293 FLIP-IN cells stably expressing His<sub>6</sub>-Ub were transfected with the indicated siRNA and treated with doxycycline (DOX) for 24 h to induce His<sub>6</sub>-Ub expression. Cells were pre-treated with 10 mM N-ethylmaleimide for 30 min and lysed in denaturing lysis buffer (6 M guanidinium-HCl, 0.1 M Na<sub>2</sub>HPO<sub>4</sub>/NaH<sub>2</sub>PO<sub>4</sub>, 10 mM Tris-HCl, 5 mM imidazole, 0.01 M  $\beta$ -mercaptoethanol, complete protease inhibitor cocktail). Lysates were sonicated on ice twice for 10 s with 1 min break and centrifuged at 15,000g for 10 min at 4°C. The supernatant was incubated with Ni-NTA-agarose beads (Qiagen) for 4 h at 4°C. Beads were collected by centrifugation, washed once with denaturing lysis buffer, once with wash buffer (8 M urea, 0.1 M Na<sub>2</sub>HPO<sub>4</sub>/NaH<sub>2</sub>PO<sub>4</sub>, 10 mM Tris-HCl, 5 mM imidazole, 0.01 M  $\beta$ -mercaptoethanol, complete protease inhibitor cocktail), and twice with wash buffer supplemented with 0.1% Triton X-100, and eluted in elution buffer (0.2 M imidazole, 0.15 M Tris-HCl, 30% glycerol, 0.72 M  $\beta$ -mercaptoethanol, 5% SDS) before analysis by SDS-PAGE and immunoblotting. **Homologous-recombination-based repair assays.** Parental U2OS cells and U2OS cells stably expressing wild-type CtIP or CtIP(T847E) mutant were transfected with the indicated siRNA and the PALB2-KR construct, synchronized with a single thymidine block, treated with doxycycline to induce CtIP expression and subsequently blocked in G1 phase by adding 40  $\mu$ M lovastatin. Cells were collected by trypsinization, washed once with PBS and electroporated with 2.5  $\mu$ g of sgRNA plasmid and 2.5  $\mu$ g of donor template using the Nucleofector technology (Lonza; protocol X-001). Cells were plated in medium supplemented with 40  $\mu$ M lovastatin and grown for 24 h before flow cytometry analysis.

**PALB2 chemical ubiquitylation.** PALB2 (1–103) polypeptides, engineered with only one cross-linkable cysteine, were ubiquitylated by cross-linking alkylation, as previously described<sup>38,39</sup>, with the following modifications. Purified PALB2 cysteine mutant (final concentration of 600  $\mu$ M) was mixed with His<sub>6</sub>-TEV-ubiquitin G76C (350  $\mu$ M) in 300 mM Tris pH 8.8, 120 mM NaCl and 5% glycerol. Tris(2-carboxyethyl)phosphine (TCEP) (Sigma-Aldrich) reducing agent was added to a final concentration of 6 mM to the mixture and incubated for 30 min at room temperature. The bi-reactive cysteine cross-linker, 1,3-dichloroacetone (Sigma-Aldrich), was dissolved in dimethylformamide and added to the protein mix to a final concentration of 5.25 mM. The reaction was allowed to proceed on ice for 1 h, before being quenched by the addition of 5 mM  $\beta$ -mercaptoethanol. His<sub>6</sub>-TEV-ubiquitin-conjugated PALB2 was enriched by passing over Ni-NTA-agarose beads (Qiagen).

**Statistics and randomization.** No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

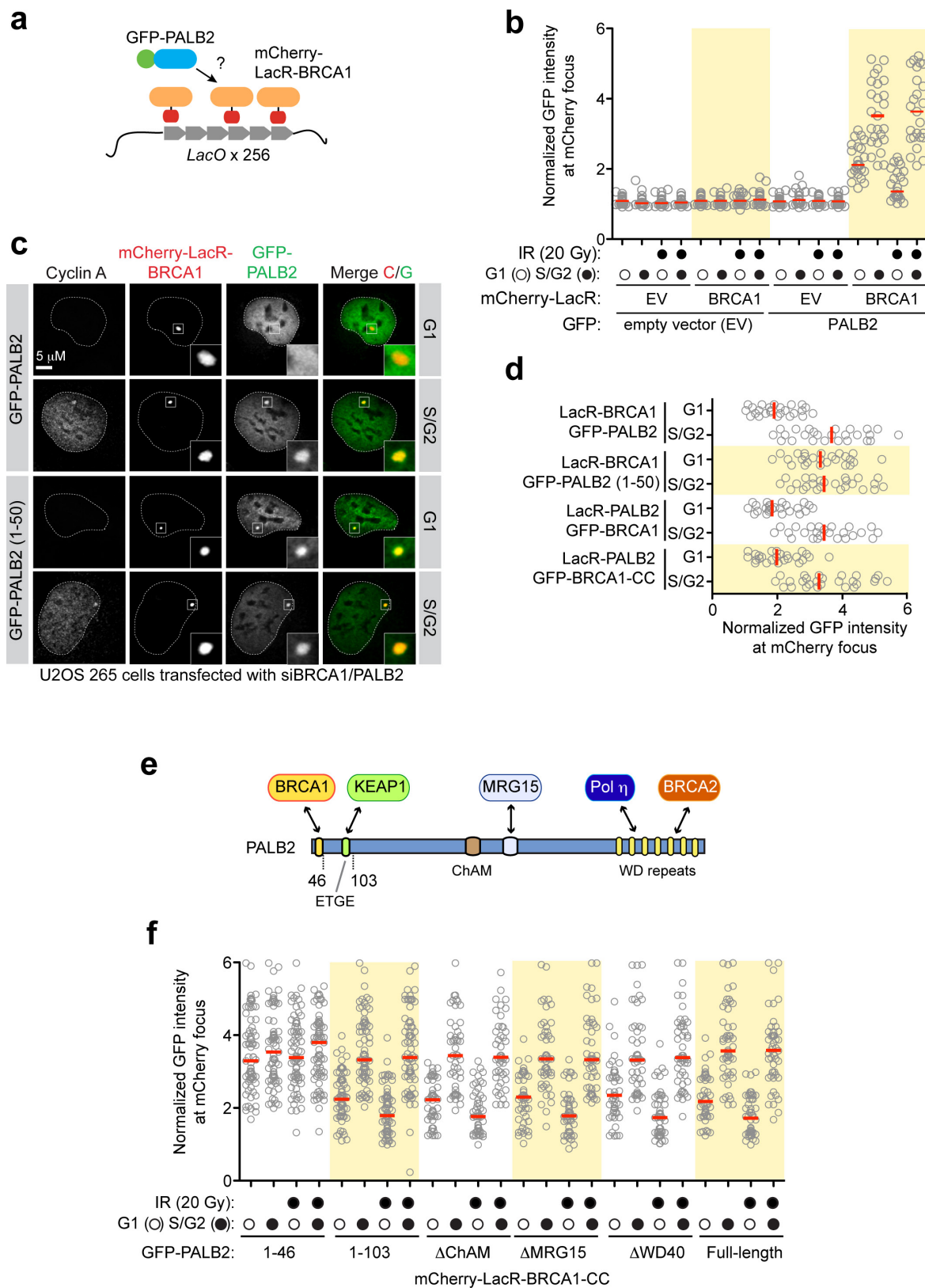
29. Fradet-Turcotte, A. *et al.* 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature* **499**, 50–54 (2013).
30. Enchev, R. I., Schreiber, A., Beuron, F. & Morris, E. P. Structural insights into the COP9 signalosome and its common architecture with the 26S proteasome lid and eIF3. *Structure* **18**, 518–527 (2010).
31. Ran, F. A. *et al.* Genome engineering using the CRISPR–Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).
32. Xia, B. *et al.* Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol. Cell* **22**, 719–729 (2006).
33. Orthwein, A. *et al.* Mitosis inhibits DNA double-strand break repair to guard against telomere fusions. *Science* **344**, 189–193 (2014).
34. Panier, S. *et al.* Tandem protein interaction modules organize the ubiquitin-dependent response to DNA double-strand breaks. *Mol. Cell* **47**, 383–395 (2012).
35. Juang, Y. C. *et al.* OTUB1 co-opts Lys48-linked ubiquitin recognition to suppress E2 enzyme function. *Mol. Cell* **45**, 384–397 (2012).
36. Cui, J. *et al.* Glutamine deamidation and dysfunction of ubiquitin/NEDD8 induced by a bacterial effector family. *Science* **329**, 1215–1218 (2010).
37. Hendriks, I. A., Schimmel, J., Eifler, K., Olsen, J. V. & Vertegaal, A. C. Ubiquitin-specific protease 11 (USP11) deubiquitinates hybrid small ubiquitin-like modifier (SUMO)-ubiquitin chains to counteract RING finger protein 4 (RNF4). *J. Biol. Chem.* **290**, 15526–15537 (2015).
38. Long, L., Furgason, M. & Yao, T. Generation of nonhydrolyzable ubiquitin-histone mimics. *Methods* **70**, 134–138 (2014).
39. Yin, L., Krantz, B., Russell, N. S., Deshpande, S. & Wilkinson, K. D. Nonhydrolyzable diubiquitin analogues are inhibitors of ubiquitin conjugation and deconjugation. *Biochemistry* **39**, 10001–10010 (2000).





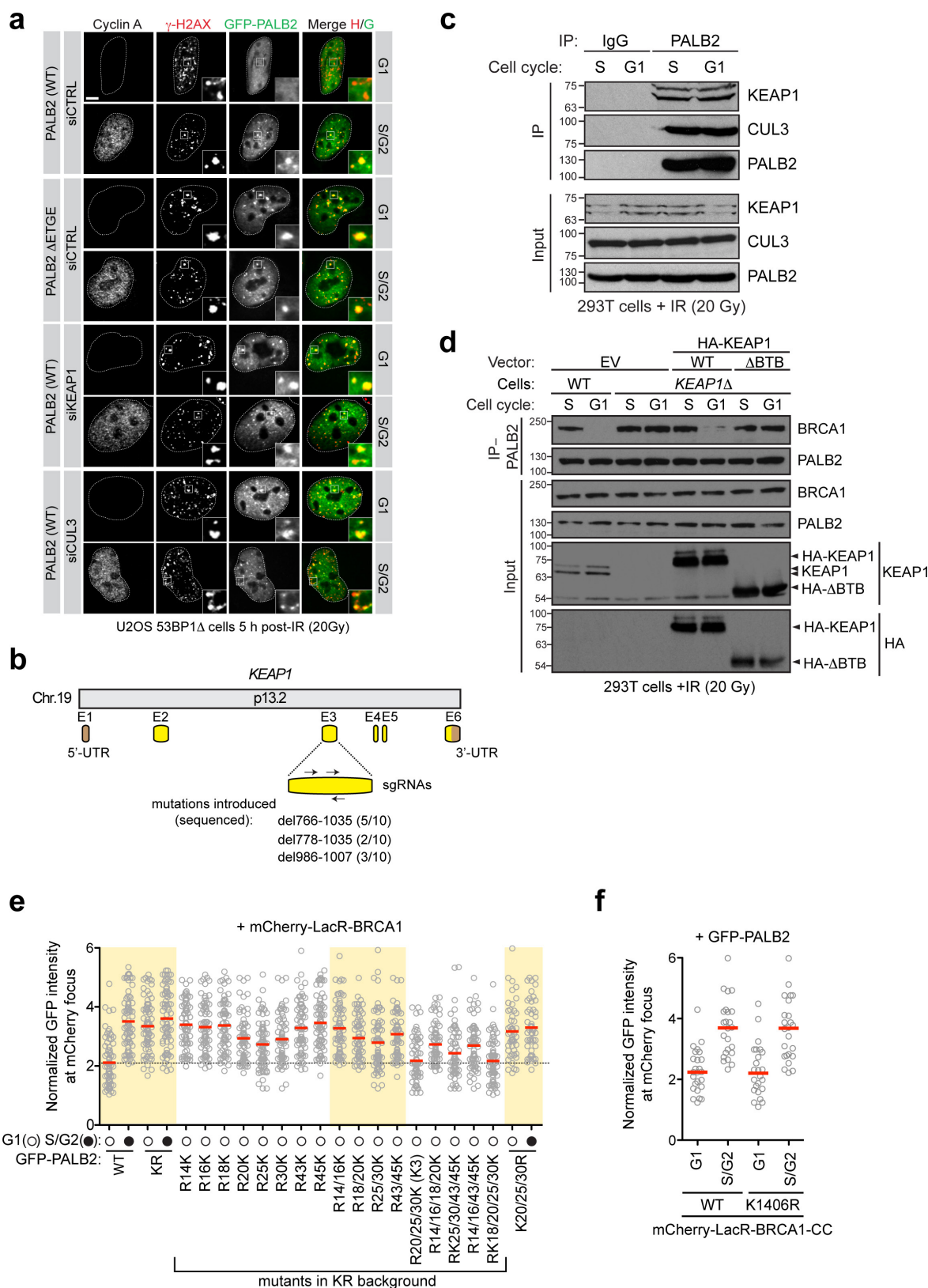
**Extended Data Figure 1 | Suppression of PALB2–BRCA2 accumulation at DSB sites in G1 53BP1Δ cells.** **a**, Schematic representation of human 53BP1 gene organization and targeting sites of sgRNAs used. Boxes indicate exons (E: yellow, coding sequence; brown, untranslated regions (UTRs)). The indels introduced by CRISPR–Cas9 and their respective frequencies are indicated. **b**, Wild-type (WT) and 53BP1Δ U2OS cells were mock- or X-irradiated (10 Gy) before being processed for 53BP1 fluorescence microscopy. DAPI was used to stain DNA and trace the outline of the nucleus. **c**, Wild-type and 53BP1Δ U2OS cells were processed for 53BP1 immunoblotting. Tubulin was used as a loading

control. **d**, Wild-type and 53BP1Δ U2OS cells either synchronized in G1 following a double-thymidine block and release or asynchronously dividing (ASN), were irradiated (2 Gy) and processed for γ-H2AX, PALB2, BRCA2 and BRCA1 immunofluorescence. The micrographs relating to BRCA1 and BRCA2 staining in G1 are found in Fig. 1a. **e**, Wild-type and 53BP1Δ U2OS cells synchronized in G1 after release from a double-thymidine block were irradiated (20 Gy) and processed for γ-H2AX, BRCA1 and BRCA2 immunofluorescence. On the left are representative micrographs for the G1-arrested cells and the quantitation of the full experiment is shown on the right (mean ± s.d.,  $N = 3$ ).



**Extended Data Figure 2 | The BRCA1–PALB2 interaction is cell cycle regulated.** **a**, Schematic of the *lacO*/LacR chromatin-targeting system. **b**, U2OS 256 cells were transfected with the indicated mCherry-LacR and GFP fusions. GFP fluorescence was measured at the site of the *lacO*-array-localized mCherry focus. Each circle represents one cell analysed and the bar is at the median. Cells were also stained with a cyclin A antibody to determine cell cycle position ( $N = 3$ ). IR, ionizing radiation. **c**, Representative micrographs of U2OS 256 cells transfected with the

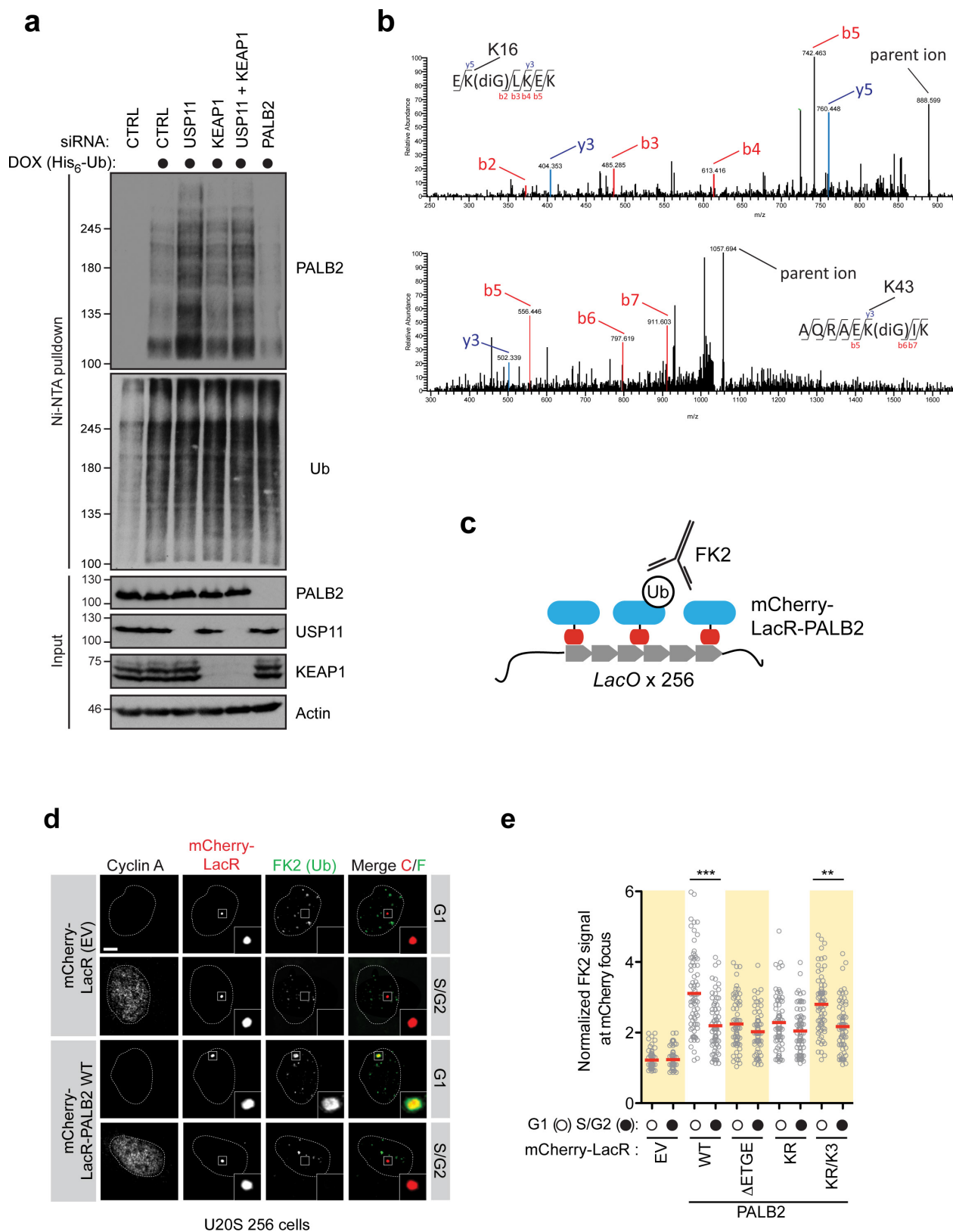
indicated mCherry-LacR and GFP fusions; data are quantified in **d**. **d**, Quantification of U2OS 256 cells transfected with the indicated mCherry-LacR and GFP fusions to tether either BRCA1 or PALB2 to the *lacO* array ( $N = 3$ ). **e**, Schematic representation of PALB2 architecture and its major interacting proteins. **f**, Quantification of U2OS 256 cells transfected with the indicated GFP–PALB2 mutants and mCherry-LacR–BRCA1-CC. Cells were also stained with a cyclin A antibody to determine cell cycle position ( $N = 3$ ).



**Extended Data Figure 3 | Inhibition of the BRCA1–PALB2 interaction in G1 depends on CUL3–KEAP1.** **a**, Representative micrographs of the experiment shown in Fig. 1d. **b**, Schematic representation of human *KEAP1* gene organization and targeting sites of sgRNAs used as described in Extended Data Fig. 1. **a**, The indels introduced by CRISPR–Cas9 and their respective frequencies are indicated. **c**, Immunoprecipitation (IP) of PALB2 from extracts prepared from irradiated 293T cells. Immunoprecipitation with normal IgG was performed as a control. **d**, 293T cells with the indicated genotypes were transfected with the indicated HA–KEAP1 constructs, synchronized in G1 or S phases and irradiated. Cells were processed for

PALB2 immunoprecipitation (IP). EV, empty vector; WT, wild type. **e**, Quantification of U2OS 256 cells transfected with the indicated GFP–PALB2 mutants and mCherry–LacR–BRCA1. Cells were also stained with a cyclin A antibody to determine cell cycle position ( $N = 3$ ). **f**, Quantification of U2OS 256 cells transfected with GFP–PALB2 and mCherry–LacR–BRCA1–CC (wild type or K1406R mutant). Cells were also stained with a cyclin A antibody to determine cell cycle position. This panel shows that the sole lysine in the PALB2–interaction motif of BRCA1 is not involved in the cell cycle regulation of the PALB2–BRCA1 interaction. **e**, **f**, Each circle represents a cell analysed and the bar is at the median ( $N = 3$ ).





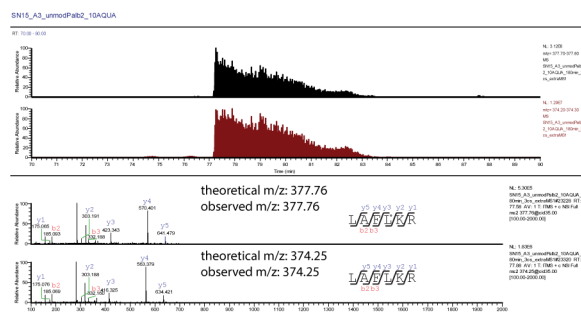
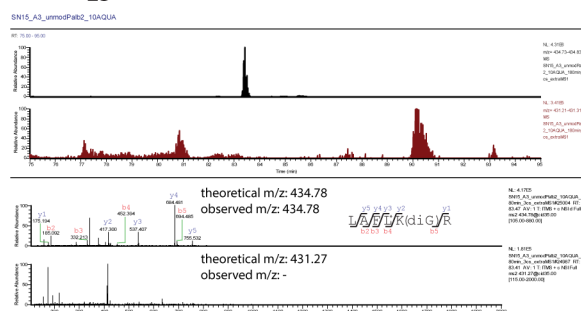
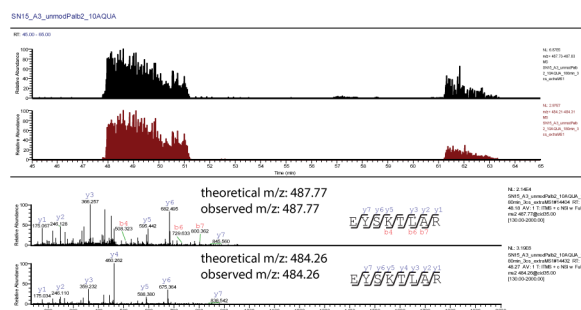
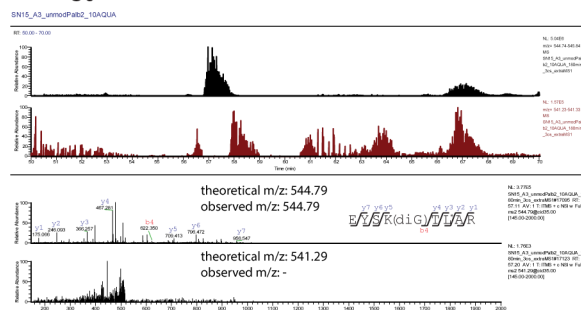
#### Extended Data Figure 4 | PALB2 is ubiquitinated by CRL3-KEAP1.

**a**, HEK293 Flp-In T-REX cells expressing doxycycline (DOX)-inducible His<sub>6</sub>-Ub were transfected with the indicated siRNAs. Cells were processed for Ni-NTA pulldown. **b**, 293T cells transfected with the indicated siRNA targeting USP11 and a Flag-PALB2 expression vector were processed for Flag immunoprecipitation followed by mass spectrometry (MS). Representative MS/MS spectra of tryptic diglycine (diG)-PALB2 peptides identified are shown (K16, top; K43, bottom). **c**, Schematic of the *lacO*/LacR chromatin-targeting system and the *in vivo* quantification of ubiquitinated PALB2.

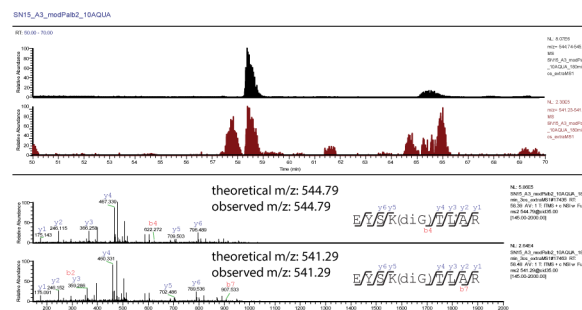
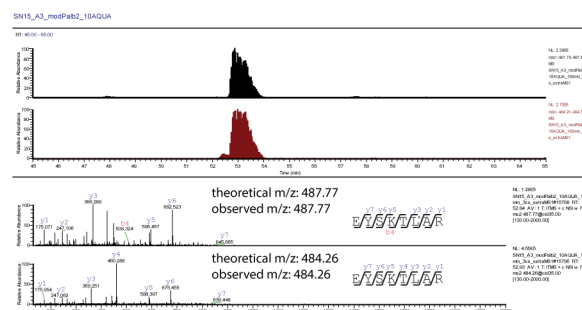
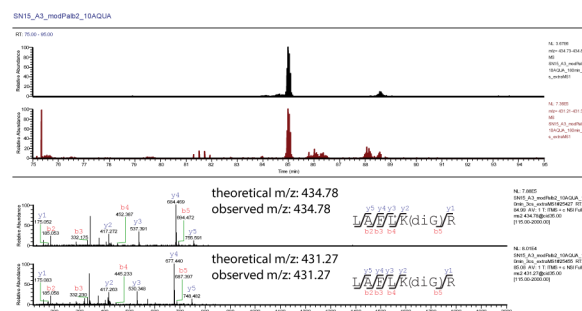
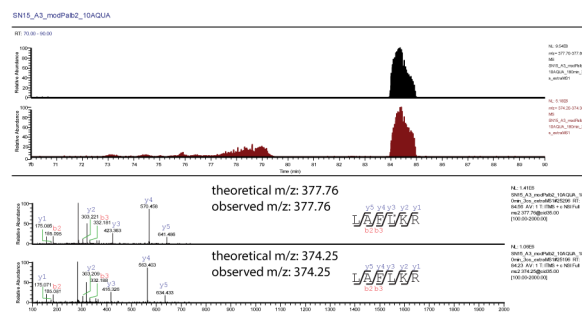
**d**, Representative micrographs of U2OS 256 cells transfected with the indicated mCherry-LacR-PALB2 vectors. Cells were processed for FK2 immunofluorescence. EV, empty vector. Scale bar, 5 μm. **e**, Quantification of U2OS 256 cells transfected with the indicated mCherry-LacR-PALB2 vectors. Cells were processed for quantification of FK2 fluorescence at the LacO focus. Each circle represents a cell analysed and the bar is at the median ( $N=3$ ). Cells were also stained with a cyclin A antibody to determine cell cycle position. Statistical significance was determined by a Kruskal-Wallis test (\*\*\* $P < 0.001$ ; \*\* $P < 0.01$ ).



## Control (-CUL3)

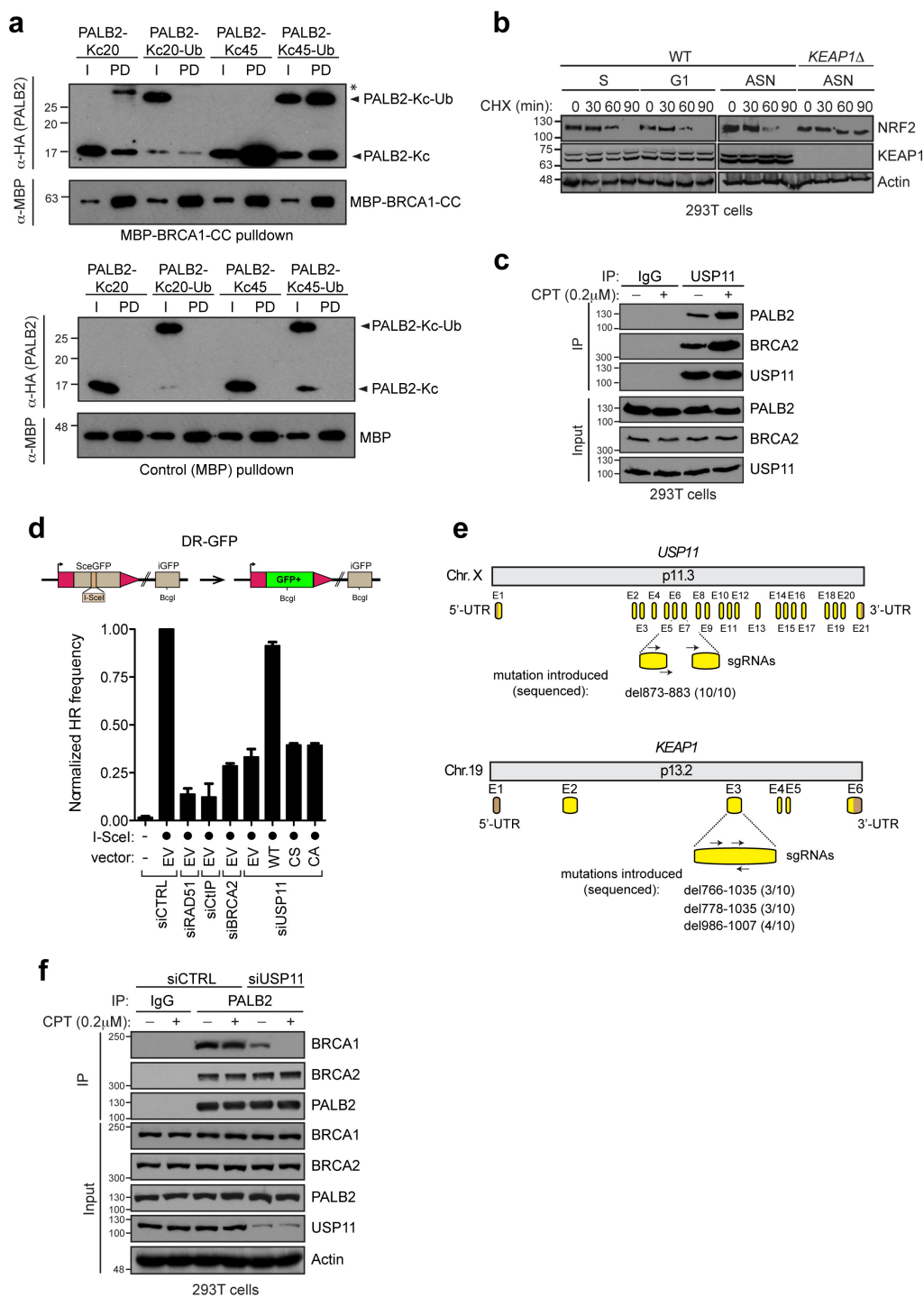
LAFLK<sub>25</sub>RLAFLK<sub>25</sub>(GG)REYSK<sub>30</sub>TLAREYSK<sub>30</sub>(GG)TLAR

## +CUL3



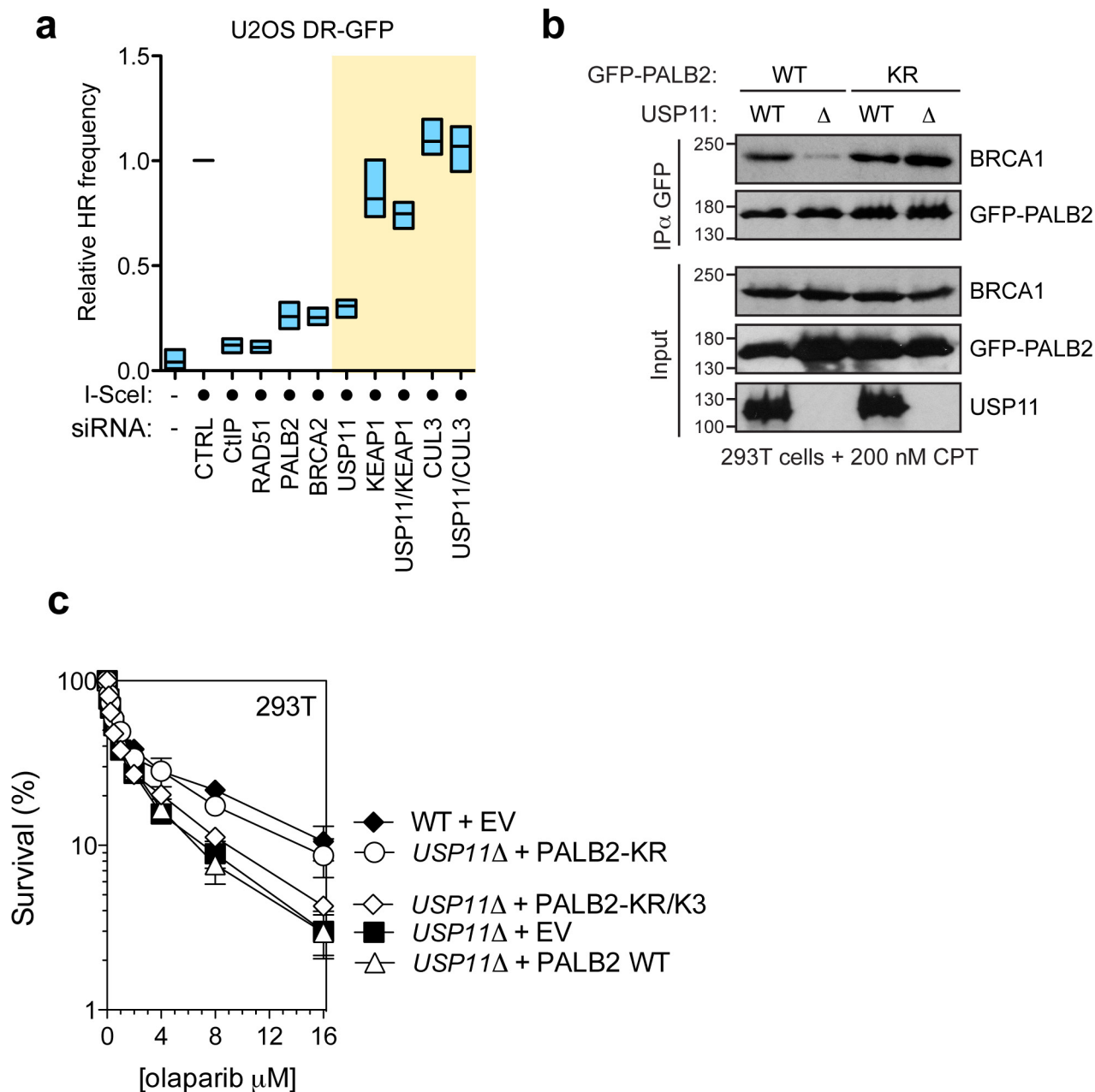
**Extended Data Figure 5 | Analysis of PALB2 ubiquitylation by mass spectrometry.** HA-PALB2 (1–103) was subjected to *in vitro* ubiquitylation reactions that lacked (left) or included (right) CUL3. Upon trypsin digestion of complete reaction products, 10 heavy labelled AQUA peptides representing N-terminal PALB2 peptides (see Methods for more information) were spiked into the peptide mixture before tandem mass spectrometry (MS/MS) analysis. Representative fragmentation spectra of AQUA peptides and unlabelled peptides from the reaction

products are shown. For each peptide, the traces from top to bottom show: mass range chromatograms (0.1 *m/z* range surrounding the *m/z* of the doubly charged peptide) of the heavy and unlabelled peptide, respectively; representative MS/MS fragmentation spectra including assigned peaks of the heavy- and light-labelled peptide, respectively. The <sup>13</sup>C<sup>15</sup>N heavy-labelled amino acid is indicated by an asterisk and the theoretical and observed *m/z* of the doubly charged peptide are indicated in the relevant spectra.



**Extended Data Figure 6 | Analysis of KEAP1- and USP11-dependent modulation of PALB2 and homologous recombination.** **a**, Site-specific chemical ubiquitylation of HA-PALB2 (1–103) at residue 20 (PALB2-Kc20-Ub) and 45 (PALB2-Kc45-Ub) was carried out by dichloroacetone linking. The resulting ubiquitylated PALB2 polypeptides along with their unmodified counterparts were subjected to pulldown with a fusion of MBP with the coiled-coil domain of BRCA1 (MBP-BRCA1-CC). I, input; PD, pulldown. Asterisk indicates a non-specific band. **b**, Wild-type and KEAP1 $\Delta$  293T cells were treated with cycloheximide (CHX) for the indicated time and then processed for NRF2 and KEAP1 immunoblotting. Actin levels were also determined as a loading control. **c**, Immunoprecipitation (IP) of USP11 from extracts prepared from 293T cells that were or were not treated with camptothecin (CPT; 200 nM). Immunoprecipitation with normal IgG was performed as a control. **d**, U2OS DR-GFP cells were transfected with the indicated siRNAs. Twenty-four hours post-transfection, cells were further transfected with

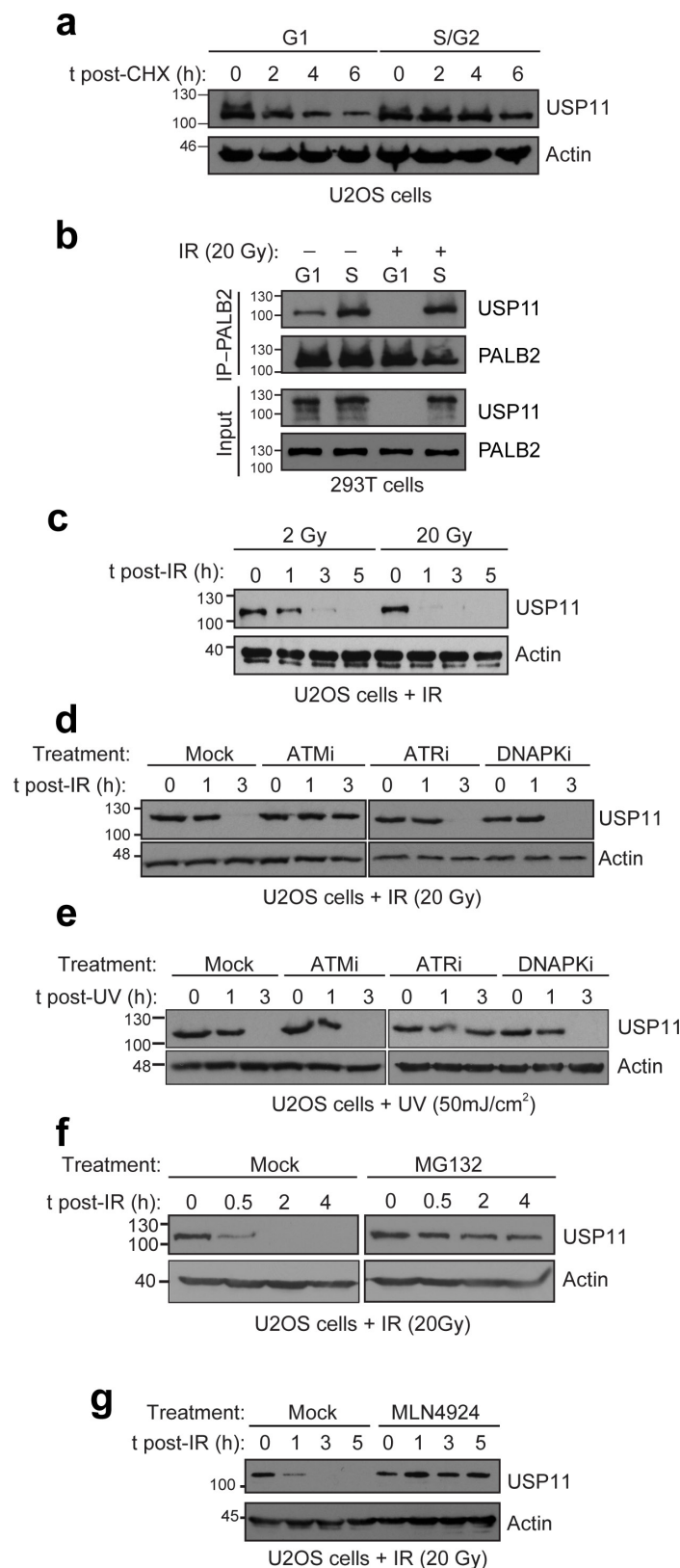
the indicated siRNA-resistant USP11 expression vectors (WT, wild type; CS, C318S and CA, C318A catalytically dead mutants) or an empty vector (EV), with or without an I-SceI expression vector. The percentage of GFP-positive cells was determined 48 h post-plasmid transfection for each condition and was normalized to the I-SceI plus non-targeting (siCTRL) condition (mean  $\pm$  s.d.,  $N = 3$ ). **e**, Schematic representation of human USP11 (top) and KEAP1 (bottom) gene organization and targeting sites of sgRNAs (as described in Extended Data Fig. 1a) used to generate the USP11 $\Delta$  and USP11 $\Delta$ /KEAP1 $\Delta$  293T cells. The indels introduced by the CRISPR-Cas9 and their respective frequencies are indicated. The USP11 knockout was created first and subsequently used to make the USP11 $\Delta$ /KEAP1 $\Delta$  double mutant. **f**, Immunoprecipitation of PALB2 from extracts prepared from 293T cells transfected with the indicated siRNA and with or without CPT (200 nM) treatment. Immunoprecipitation with normal IgG was performed as a control.



**Extended Data Figure 7 | USP11 antagonizes KEAP1 action on PALB2.**

**a**, U2OS DR-GFP cells were transfected with the indicated siRNAs or left untransfected (–). Twenty-four hours post-transfection, cells were transfected with an I-SceI expression vector (circle). The percentage of GFP-positive cells was determined 48 h post-plasmid transfection for each condition and was normalized to the I-SceI plus non-targeting (CTRL) condition (mean  $\pm$  range,  $N = 3$ ). **b**, Parental 293T cells (wild type (WT))

or a USP11 $\Delta$  derivative were transfected with the indicated GFP-PALB2 constructs, treated with CPT and processed for GFP immunoprecipitation (IP). **c**, Parental 293T cells (wild type) or a USP11 $\Delta$  derivative were transfected with an empty vector (EV) or the indicated PALB2 expression vectors. Sensitivity of the cells to the PARP inhibitor olaparib was then determined by a clonogenic survival

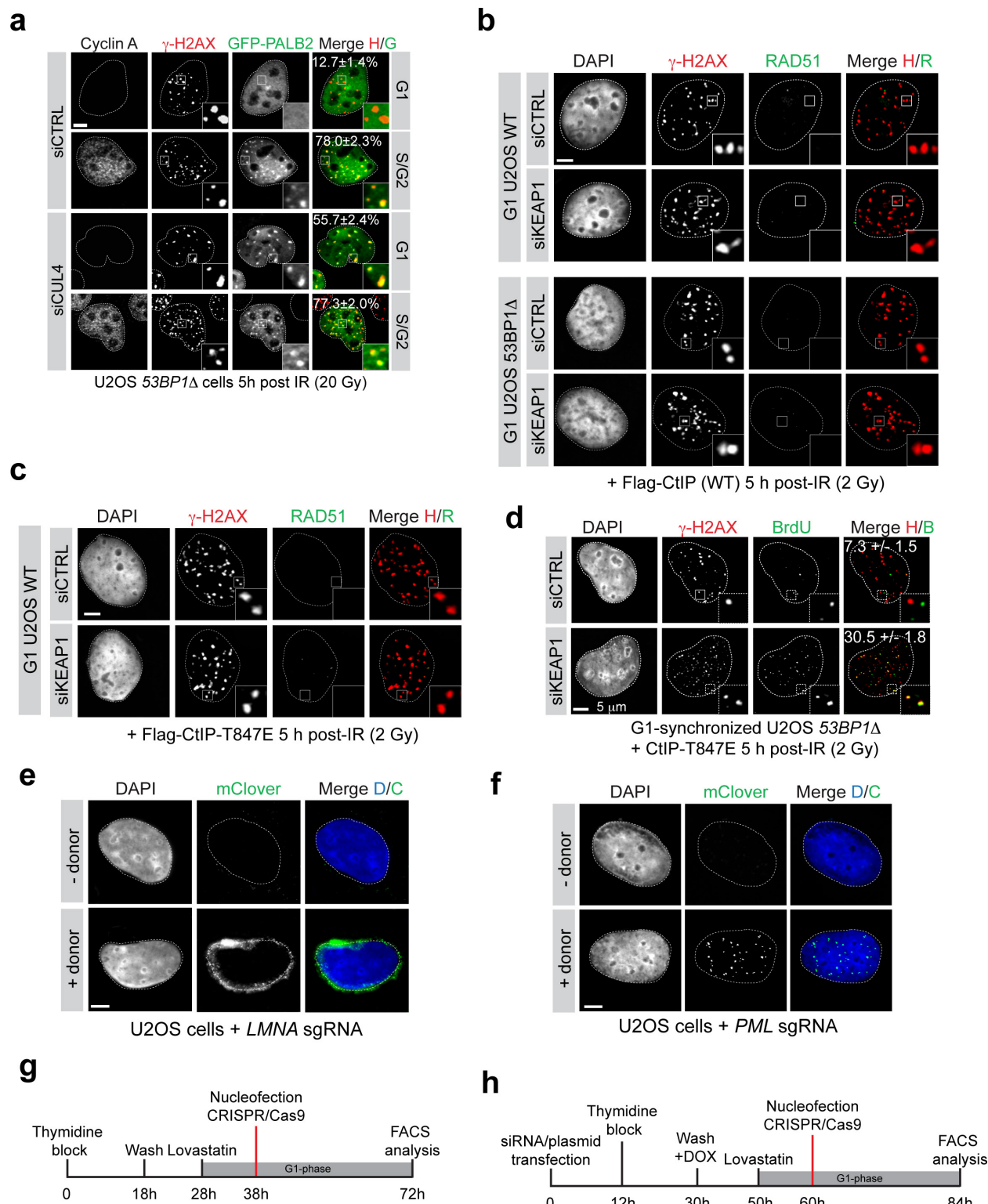


#### Extended Data Figure 8 | Characterization of USP11 protein stability.

**a**, U2OS cells synchronized in G1 or S/G2 were treated with cyclohexamide (CHX) and processed at the indicated time points to monitor USP11 stability. **b**, Immunoprecipitation (IP) of PALB2 from extracts prepared from 293T cells that were synchronized in G1 or S phase and treated or not with ionizing radiation (IR; 20 Gy). **c**, U2OS cells were irradiated with a dose of 2 or 20 Gy and processed for USP11 immunoblotting at the indicated times post-ionizing radiation. Actin was used as a loading control. **d**, U2OS cells, mock treated or incubated with the ATM inhibitor KU55933 (ATMi), ATR

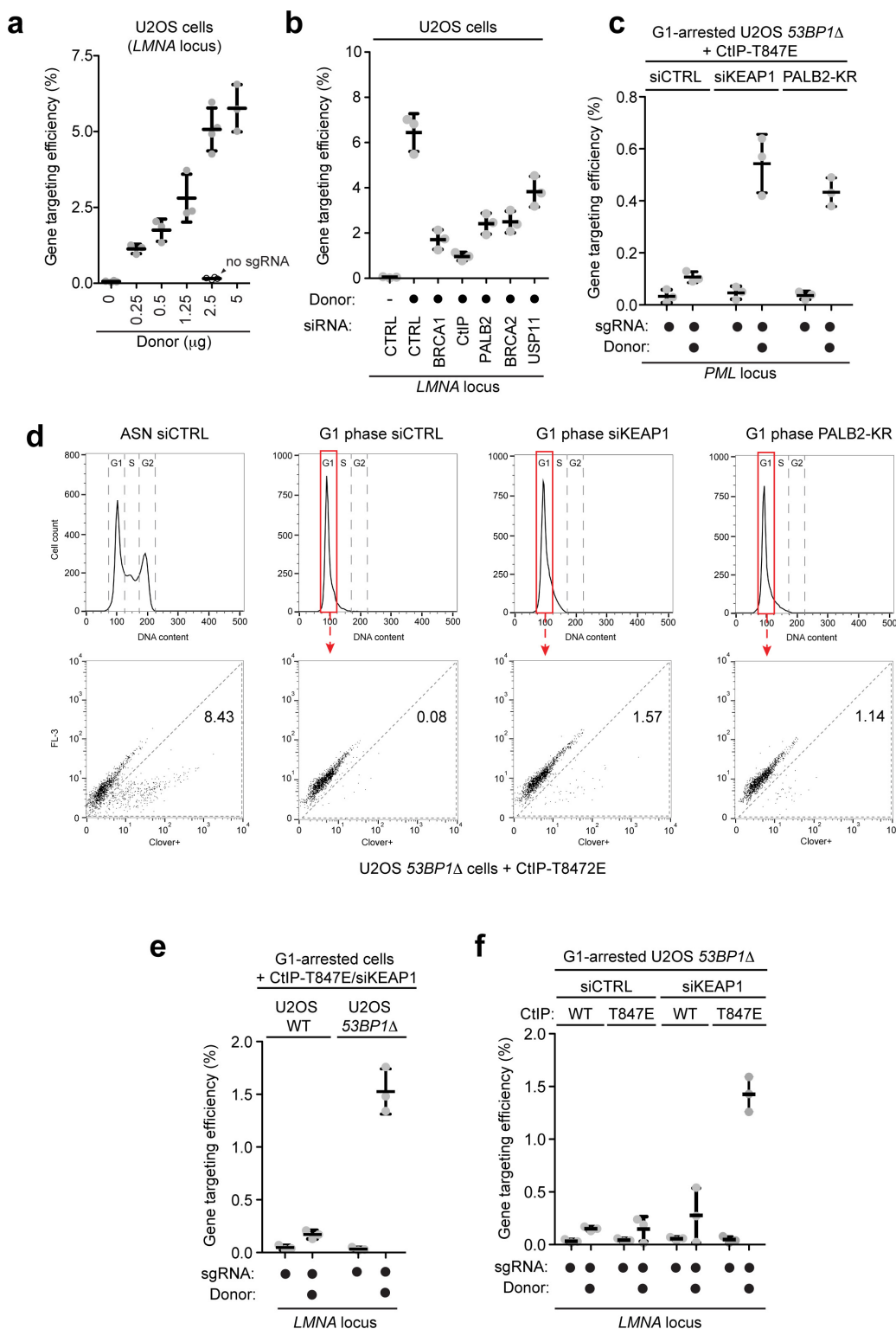
inhibitor VE-821 (ATRi) or DNA-PKcs inhibitor NU7441 (DNAPKi), were irradiated (20 Gy) and processed for USP11 and actin (loading control) immunoblotting. **e**, Similar experiment to **d** except that cells were exposed to ultraviolet (UV) radiation (50 mJ cm<sup>-2</sup>). **f**, U2OS cells, mock treated or incubated with the proteasome inhibitor MG132, were irradiated (20 Gy) and processed for USP11 and actin (loading control) immunoblotting. **g**, U2OS cells, mock-treated or incubated with the cullin inhibitor MLN4924, were irradiated (20 Gy) and processed for USP11 and actin (loading control) immunoblotting.





**Extended Data Figure 9 | Reactivation of RAD51 loading and unscheduled DNA synthesis in G1.** **a**, 53BP1 $\Delta$  U2OS cells were transfected with the indicated siRNA, synchronized in G1 or S/G2 by release from a double-thymidine block and irradiated (20 Gy) before being processed for fluorescence microscopy. DAPI was used to trace the nuclear boundary and cyclin A staining was used to determine cell cycle position. The percentage of cells with more than five  $\gamma$ -H2AX-colocalizing PALB2 foci is indicated as the mean  $\pm$  s.d.,  $N = 3$ . Scale bar, 5  $\mu$ m. **b**, Representative micrographs of irradiated G1-synchronized wild-type (WT) and 53BP1 $\Delta$  U2OS cells transfected with the indicated siRNA and expressing wild-type CtIP. **c**, Representative micrographs of irradiated G1-synchronized wild-type U2OS cells transfected with the

indicated siRNA and expressing CtIP(T847E). **d**, U2OS 53BP1 $\Delta$  cells were synchronized in G1, supplemented with BrdU, irradiated (2 Gy) and processed for  $\gamma$ -H2AX and BrdU immunofluorescence. The percentage of cells with more than five  $\gamma$ -H2AX-colocalizing BrdU foci is indicated (mean  $\pm$  s.d.,  $N = 3$ ). **e**, Micrograph of a U2OS cell targeted with the CRISPR-mClover system showing the typical perinuclear expression pattern of lamin A. **f**, Micrograph of a U2OS cell targeted with the mClover system showing an expression pattern characteristic of subnuclear PML foci. **g**, Timeline of the gene-targeting (LMNA) experiment presented in Fig. 4d. **h**, Timeline of the gene-targeting (LMNA or PML) experiment presented in Fig. 4e and Extended Data Fig. 10.



**Extended Data Figure 10 | Analysis of homologous recombination in G1.** **a**, Quantitation of gene targeting efficiency at the *LMNA* locus in asynchronously dividing U2OS cells transfected with increasing amount of donor template and with (grey) or without (white) sgRNAs. Gene-targeting events were detected by flow cytometry (mean  $\pm$  s.d.,  $N \geq 3$ ). **b**, Quantitation of gene-targeting efficiency at the *LMNA* locus in asynchronously dividing cells transfected with the indicated siRNA. Gene-targeting events were detected by flow cytometry (mean  $\pm$  s.d.,  $N = 3$ ). **c**, Gene-targeting efficiency at the *PML* locus measured by flow cytometry in G1-arrested *53BP1* $\Delta$  U2OS cells expressing the CtIP(T847E) mutant and co-transfected with the indicated siRNA or a PALB2-KR expression construct (mean  $\pm$  s.d.,  $N = 3$ ). **d**, Representative FACS

profiles showing the gating for 1N DNA content cells and the detection of mClover-positive cells in the *LMNA* gene targeting assay in asynchronous (ASN) or G1-arrested *53BP1* $\Delta$  U2OS cells expressing the CtIP(T847E) mutant and co-transfected with the indicated siRNA or a PALB2-KR expression construct. **e**, Gene-targeting efficiency at the *LMNA* locus measured by flow cytometry in G1-arrested parental (wild-type (WT)) and *53BP1* $\Delta$  U2OS cells transfected with KEAP1 siRNA and expressing the CtIP(T847E) mutant (mean  $\pm$  s.d.,  $N = 3$ ). **f**, Gene-targeting efficiency at the *LMNA* locus measured by flow cytometry in G1-arrested parental (wild-type) and *53BP1* $\Delta$  U2OS cells transfected with the indicated siRNA and expressing either wild type or the CtIP(T847E) mutant (mean  $\pm$  s.d.,  $N = 3$ ).

# CAREERS

**TRAINING** Industry-supported PhDs offer practical advantages [go.nature.com/xtgamf](http://go.nature.com/xtgamf)

**COMMUNICATION** Tips for crafting a good elevator pitch [go.nature.com/m9bwv4](http://go.nature.com/m9bwv4)

**NATUREJOBS** For the latest career listings and advice [www.naturejobs.com](http://www.naturejobs.com)



## MENTORSHIP

# Stewards of China's future

*The 2015 Nature Awards for Mentoring in Science recognize Chinese scientists who have invested in the next generation.*

BY ED GERSTNER

During the past two decades, Chinese science has undergone profound growth. China's investment in research and development surpassed that of the European Union in 2013, and it is predicted to overtake that of the United States by the end of the decade (see *Nature* <http://doi.org/w5r>; 2014). The proportion of published scientific papers that include Chinese co-authors has jumped from 2.4% in 1997 to 19% in 2014 — second only to the US contribution last year of 25%.

Those statistics are impressive. But if China

is to become a true scientific superpower, it must be able to produce great scientists who are not just knowledgeable but also creative and skilled in innovation. And great scientists need great mentors to lead the way.

In recognition of the vision, dedication and hard work of those charged with nurturing the next generation of Chinese researchers, this year's *Nature* Awards for Mentoring in Science honour five researchers in China. The winners, feted in an 8 December ceremony, were chosen by panels composed of Chinese scientists and Springer Nature editorial representatives (see [go.nature.com/hdi5k7](http://go.nature.com/hdi5k7)). Submissions included

statements from five people who had been mentored by the nominee and statements from the nominees reflecting their own thoughts on mentoring.

Owing to China's size, submissions were divided into 'north' and 'south', with awards for lifetime and mid-career achievement in each. The 50,000-yuan (US\$7,815) lifetime-achievement award for northern China was shared between immunologist Xuetao Cao, who is president of the Chinese Academy of Medical Sciences, and plant scientist Xingwang Deng, dean of the School of Advanced Agricultural Sciences at Peking University. The winner for southern China is Hongyuan Chen, an electro-analytical chemist and director of the Institute of Chemical Biology at Nanjing University.

In the mid-career category, the 50,000-yuan awards for northern and southern China went, respectively, to Yigong Shi, a structural biologist and dean of life sciences at Tsinghua University in Beijing, and Hongbing Shu, an immunologist at Wuhan University.

## CHALLENGE TO CONVENTION

Like many Asian nations, China is often seen as a place of rigid hierarchies rooted in deference to power. One trait shared by all the winners, and indeed by all those nominated, is an understanding that the only authority in science is evidence — and that conventional wisdom must always be open to question.

Shi, who was named a chair professor of molecular biology at Princeton University in New Jersey before he returned to China in 2008, thinks that most Chinese students are too wary of contradicting senior researchers and accepted scientific ideas. "I encourage my students to think critically and to challenge the authorities, including myself, so that they can learn that established rules can be broken, and with that, new fields of research can be built," he says.

Cao agrees. "We should inspire students to have confidence to challenge the dogma in the textbook and address fundamental questions in science," he says.

The lesson is not lost on the winners' protégés. "The scientific literature is a baffling mass of conflicting ideas and results, accepted wisdom and false assumptions," notes Weilin Chen, a cancer immunologist at Zhejiang University and one of Cao's former PhD students at the Second Military Medical University in Shanghai. "Professor Cao often said that creativity comes from different directions with different views," she says. "And he treats ►



► everyone, regardless of whether they are a PhD student or a visiting scholar, with the same high regard.”

In the past, most Chinese labs were indeed quite rigid, with a single senior professor directing junior professors, postdocs and students along strictly hierarchical lines. With the rapid expansion of research institutes, however — fuelled by a large influx of researchers returning from overseas — the structure of many labs has begun to follow a less-hierarchical model, with many independent principal investigators all pursuing their own agendas and research directions.

### A TEACHER'S PHILOSOPHY

The mentors honoured by *Nature* have recognized the importance of instilling young researchers with the self-confidence that they need to establish their own intellectual identity and to make their own way in the world. “In my opinion, simply imparting knowledge is not enough,” says Hongyuan Chen. “A mentor should teach students the way of thinking. In the area of science, I guide my students to think in a scientific way, and give them the opportunity to solve problems independently.”

He thinks that a good mentor must have a keen sense of when a student requires guidance and when he or she needs freedom. “For students who are just starting out, we need to give them more-detailed instructions to let them get used to research gradually,” he says. “And for those who have a solid knowledge base, strong independence and creativity, I let them think and practise in their own ways.”

Jingjuan Xu, a former PhD student of Hongyuan Chen's and now an analytical chemist at Nanjing University, says that Chen provided an open environment that fostered imagination and creativity. “He encouraged us to read philosophy and literature, and think from different aspects,” recalls the chemist. “He said that every student is an independent, thinking being; a good mentor should nurture them to become ‘horses’ rather than ‘sheep.’”

Good mentors also recognize that it is not enough to produce successful scientists — it is just as important to teach others how to be effective, inspiring leaders themselves. Lei Li, a postdoc of Deng's at Yale University in New Haven, Connecticut, and now a professor in the School of Life Sciences at Peking University, recounts her own training in Deng's lab. “As I became more senior in the lab, Professor Deng started to ask me to help others in their lab techniques and in reading their manuscripts, which I soon realized was part of a system,” she says. “When he discovered performance issues, he never just criticized; he took time to find the root of the problem. And in several instances,

**“We should inspire students to have confidence to challenge the dogma in the textbook.”**



Immunologist Xuetao Cao (left) and plant scientist Xingwang Deng (right) both won mentoring awards.

he delegated me to do the pep talk.”

The testimonials for the award winners all strongly reflect the scientists' unwavering dedication to the success of their protégés. But one story in particular stands out.

In 2005, immunologist Bo Zhong, now at Wuhan University, applied to do a PhD in Hongbing Shu's lab after graduating with a major in English. “I was determined to study biology after graduation because I was interested in nature,” says Zhong. At Wuhan, “Dr Shu had recently been appointed as dean of life sciences, and his group [at the National Jewish Medical and Research Center in Denver, Colorado] had just published a milestone discovery in *Molecular Cell*. Every student with ambition wanted to join his lab — and so did I”.

### NEVER GIVE UP

Zhong knew that it wouldn't be easy. “I had to admit that my background was much weaker than those who majored in biology,” he says. “I downloaded all his publications but found that I could hardly understand them. I knocked on the door to his office, and asked many naive questions. He patiently explained the details, recommended more publications to me and encouraged me to ask him if I had any difficulty in understanding the studies. Following his instructions, I read more papers, and wrote a five-page summary about pattern recognition and signalling, and asked whether I could join his lab. To my surprise, he agreed.”

Shu admits that he was unsure about Zhong's potential at first, but after seeing his determination, Shu felt that Zhong deserved a chance to show what he could do. He doesn't regret the decision. “After I was convinced of his ambition and drive for a scientific career, I took him without hesitation. He has so far proved himself as one of the most successful students trained in my lab.” After taking him on, Shu asked Zhong to turn the summary that he had written into a full review paper, which



RIGHT: YALE UNIV.

became the first publication to come out of the newly formed lab.

Shu thinks that patience and perseverance are among the most important traits of good mentorship, something he learnt from one of his own mentors: his PhD supervisor, Harish Joshi, a cell biologist at Emory University in Atlanta, Georgia. “I have always remembered what he told me when I was in his lab. ‘Do not fire them; fire them up!’” Shu recalls. “In my 17- years' mentoring life, I have never given up on any one of my students.”

A well-known Chinese saying goes, “If someone is your teacher for just one day, you should regard that person as your parent for the rest of your life.” The influence that great mentors have does indeed live long — and not just in their students, but in their students' students. “When I started my own lab in 2012, I often asked myself what Yigong would do,” says Liang Feng, a structural biologist at Stanford University in California and a former PhD student of Shi's. “I kept all e-mail communications Yigong sent to me or to the lab, and often went back to read them. They are like a ‘how-to’ guide for running a lab. For me and many others, Yigong was not only a great mentor and a role model, but also a relentless supporter and a lifelong friend.”

The word used to describe the most revered teachers, *shifu* — a portmanteau of the words for teacher, *laoshi*, and father, *fuqin* — echoes the deep connection that forms between exceptional mentors and their protégés. None of the scientists who nominated their mentors for an award takes this filial bond for granted. In the words of Hongyuan Chen's protégé Jingjuan Xu, “I think that ‘father’ is really too high a standard to expect from a teacher. But we are the lucky children, because Professor Chen treated us like his own kids.” ■

**Ed Gerstner** is executive editor for *Nature journals in Greater China*, based in Shanghai.



# CITADEL

*How to survive the solstice.*

BY JOHN GILBEY

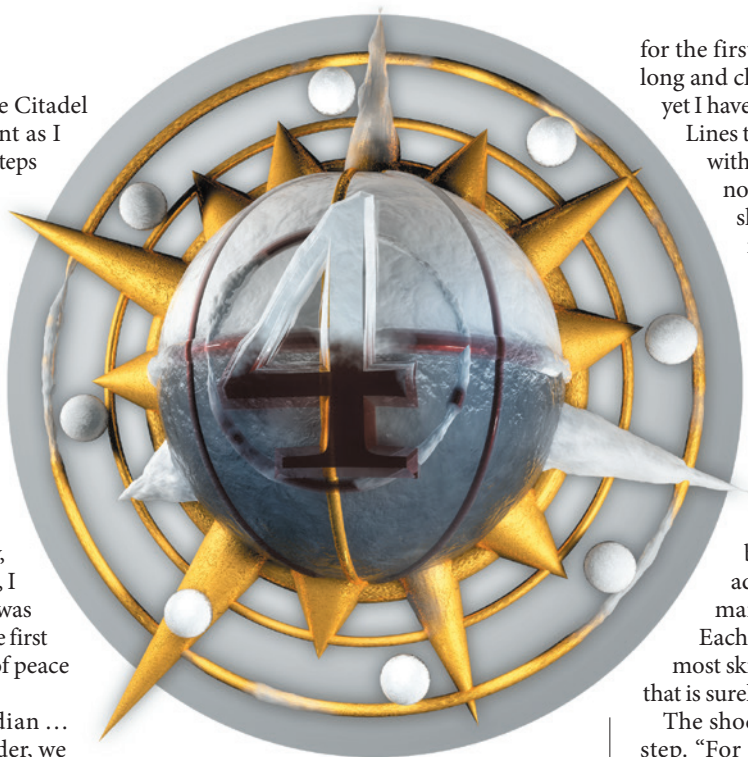
The guard at the gate of the Citadel nodded acknowledgement as I walked up the worn stone steps towards her, my boots crackling in the frost. “A cold night, pilgrim,” she intoned, her breath visible on the night air. I smiled to myself, realizing that she had mistaken my hastily snatched robes for the garb of a common supplicant. The temptation was too great and I twitched aside my cloak to reveal my cryptogram of office.

The reaction of horror was immediate, thus I decided to spare her rank — and so, possibly, her life. “Your pardon, Eminence, I took you for another...” Her fear was real and I felt a tang of guilt like the first nip of frostbite. I made the sign of peace and she relaxed, fractionally.

“We are all pilgrims, Guardian ... Worker or healer, hunter or feeder, we must all play out our duty to the faith with honour and respect according to our trade.” Aware that I was almost quoting from the creed she cast her eyes downward and muttered a blessing under her breath. I paused while she recovered herself, gauging her adherence to protocol as she scanned my cypher and released the portal. “Pass, my Lord...”

The great hall beyond the robing room was quiet, aside from the distant sound of the wind moving around the tower above. The hall had set itself for night, so only a low glow followed my progress towards the display. My Lady stood, apparently deep in reverie, before the map that once purported to describe our world — but which we now know is wildly untruthful. She looked up at my approach and pointed to a small red stain on the curved stone of the panel. “Another is gone — that makes three since midsummer...”

News indeed, and of an import that explained the urgent summons. When we first inherited this duty, only a handful of the myriad marks on the plot were red — the great majority being green, the colour of plants and life. Now, 50 summers later, close to half have adopted the deep crimson of mourning. Lore tells us that these marks speak of the health enjoyed by those sentinel obelisks of impervious metal that gird our



lands, that somehow talk to the citadel and help to build the intricate coloured patterns that scatter the map.

These red marks prey on my mind, seeming to signify the loss of so much more. So many deaths! Folk whose faces glide before me every time I visit this chamber, yet I must remain resolute — as so many others have done before me. If only my parents had not been among those who perished in that sudden, crippling spring ice-fall when I was still so small a child. The knowledge and learning that died with them cannot be replaced, yet the common people still look to the Lady and I for counsel and guidance.

As so many times before, I walked forward and laid my hands flat on the stone panels as if to commune with the hidden forces within. With a finger tip I gently traced the outlines of the glyphs that I feel certain hold the key to the secret knowledge — then hung my head in shame and frustration. I felt a warm hand slide into mine and squeeze it lightly.

“You are troubled, husband.” The smile of the Lady should be able to melt the ice fields that surround us, but that would mean sharing it with others. I turned to her, and we embraced

for the first time in many days. I held her long and close. “Our world may be dying, yet I haven’t the wisdom to read the signs.

Lines traverse the plot, growing longer with every passing solstice — yet I’ve not the skill to know whether they should alarm or reassure by their rise and fall.”

My Lady took my arm and turned me to face the empty hall. “Have you considered, husband, why the ancients who built this place made the hall so large? Surely it is not meant just for the two of us — who alone may enter it today? Perhaps one of the pilgrims who freezes in the courtyard beyond has a shred of wisdom to add to our own? Indeed, mayhap many of them have much to give?

Each traveller who is sent here is the most skilled of each hearth and trade — that is surely a sign to us...”

The shock in my face drove her back a step. “For many generations it has been thus,” I stormed. “They stand below and we assure them that all is well, that we maintain control. We alone must hold this place — else who in the land will know their point and worth?”

Her eyes, wide and dark in the poor light, clouded with anger and regret. For the second time this night I had demeaned and cast down one of my folk whose only failing was to share their humanity. Turning abruptly from her, I faced the centre of the display and made great play of deep thought. It was obvious that she was right, there could be only one conclusion.

“Very well. When the winter solstice celebration comes we will open the hall to all and demand — yes, demand — that they share their skills so that we may deepen our understanding of this place.” The ghost of a smile flitted across my Lady’s lips, before she lowered her gaze.

At the base of the plot, the image written thus “Survival Likelihood 3” in the rock became faint for a moment, and the last glyph in the sequence changed to “4”. One day, our children may understand whether that was a good thing or not. ■

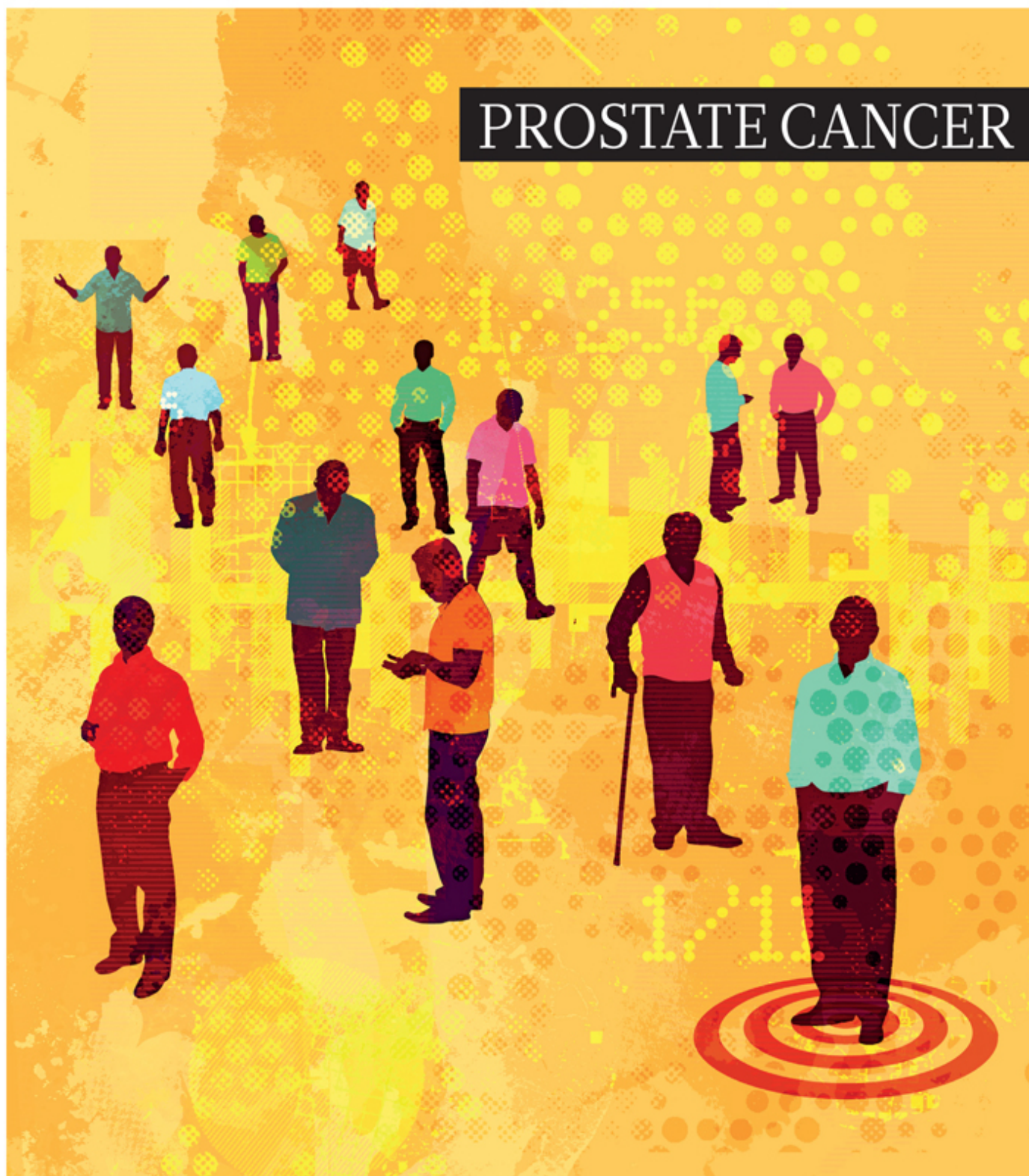
John Gilbey writes from the academic seclusion of the University of Rural England, where they worry about things like this. He tweets as @John\_Gilbey.

ILLUSTRATION BY JACEY

ON NATURE.COM  
Follow Futures:  
@NatureFutures  
go.nature.com/mtoodm

# natureOUTLOOK

## PROSTATE CANCER



Produced with support from  
Ferring Pharmaceuticals and a grant from  
Astellas Pharma Global Development, Inc. and  
Medivation, Inc.

Old aim,  
new targets



# natureOUTLOOK

## PROSTATE CANCER

17 December 2015 / Vol 528 / Issue No 7582



Cover art: Gary Neill

### Editorial

Herb Brody  
Michelle Grayson  
Richard Hodson  
Jenny Rooke

### Art & Design

Wesley Fernandes  
Mohamed Ashour  
Andrea Duffy

### Production

Karl Smart  
Ian Pope  
Mira Loutfi

### Sponsorship

Janice Stevenson  
Samantha Morley

### Marketing

Hannah Phipps

### Project Manager

Anastasia Panoutsou

### Art Director

Kelly Buckheit Krause

### Publisher

Richard Hughes

### Chief Magazine Editor

Rosie Mestel

### Editor-in-Chief

Philip Campbell

**L**ive long enough, and most men will develop prostate cancer. Globally, it is the second most common cancer in men, and in some places it takes the top spot (page S118).

As the prime reproductive years fade, the gland typically begins to misbehave. The first sign that men often experience is inflammation — a condition that is sometimes, but not always, a precursor to cancer. The interplay between inflammation and cancer remains an area of intense research (page S130).

Prostate-cancer screening has provoked contentious debate (page S120). Blood tests for prostate-specific antigen (PSA) have led to the discovery of cancers at earlier and more treatable stages. But they have also revealed many tumours that could safely be left untreated. Researchers are looking beyond PSA to other biomarkers that could be used to tell more reliably which cancers need treatment (page S124). Often, the best therapeutic option is just to be vigilant — ‘active surveillance’ is now the norm (page S126). When a trip to the operating theatre is unavoidable, robotics is making prostate surgery less likely to cause adverse effects (page S132).

Hopes for a vaccine have dimmed (page S134). The only approved immunotherapy for prostate cancer — sipuleucel-T — adds mere months to survival time and is expensive.

Researchers are focusing on combinations of therapies, such as a checkpoint therapy administered together with a drug that targets tumour hypoxia. Because prostate tumours are most dangerous once they escape the gland itself, intense efforts are targeting metastatic cancers that have become resistant to standard treatments (page S128).

We are pleased to acknowledge support from Ferring Pharmaceuticals and a grant from Astellas Pharma Global Development, Inc. and Medivation, Inc. in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

**Herb Brody**  
*Supplements Editor*

## CONTENTS

### S118 PROSTATE CANCER

#### Small organ, big reach

An outline of the common cancer

### S120 SCREENING

#### Diagnostic dilemma

Reaching a consensus on screening

### S123 PERSPECTIVE

#### Enforce the clinical guidelines

Monique Roobol discusses PSA testing

### S124 PROGNOSIS

#### Proportionate response

Identifying the truly risky cancers

### S126 TREATMENT

#### When less is more

The rise of active surveillance

### S128 METASTASIS

#### Resistance fighters

Treatments for metastatic disease

### S130 MICROBIOLOGY

#### Inflammatory evidence

The response could be a cancer cause

### S132 Q&A

#### A robot convert

Declan Murphy on robotic surgery

### S134 THERAPY

#### An immune one-two punch

Combination therapy for prostate cancer

### S137 PROSTATE CANCER

#### 4 big questions

Key areas of research

## RELATED ARTICLES

**S138 Targeting the MLL complex in castration-resistant prostate cancer**  
*R. Malik et al.*

**S147 Is docetaxel the ‘black widow’ of mCRPC drugs?**  
*B. C. Liaw & W. K. Oh*

**S149 Comprehensive validation of published immunohistochemical prognostic biomarkers of prostate cancer—what has gone wrong?**  
*F. Huber et al.*

**S158 Optimizing prostate cancer survivorship care**  
*M. J. Resnick*

**S160 Who dies from prostate cancer?**  
*A. Patrikidou et al.*

*Nature Outlooks* are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the *Nature Outlook* Editorial guidelines available at [go.nature.com/e4dwz](http://go.nature.com/e4dwz)

#### CITING THE OUTLOOK

Cite as a supplement to *Nature*, for example, *Nature* Vol. XXX, No. XXXX Suppl., Sxx–Sxx (2015).

#### VISIT THE OUTLOOK ONLINE

The *Nature Outlook Prostate Cancer* supplement can be found at <http://www.nature.com/nature/outlook/prostate-cancer>. It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

#### SUBSCRIPTIONS AND CUSTOMER SERVICES

For UK/Europe: Nature Publishing Group, Subscriptions, Brunel Road, Basingstoke, Hants, RG21 6XS, UK. Tel: +44 (0) 1256 329242. Subscriptions and customer services for Americas – including Canada, Latin America and the Caribbean: Nature Publishing Group, 75 Varick St, 9th floor, New York, NY 10013-1917, USA. Tel: +1 866 363 7860 (US/Canada) or +1 212 726 9223 (outside US/Canada). Japan/China/Korea: Nature Publishing Group – Asia-Pacific, Chiyoda Building 5-6th Floor, 2-37 Ichigaya Tamachi, Shinjuku-ku, Tokyo, 162-0843, Japan. Tel: +81 3 3267 8751.

#### CUSTOMER SERVICES

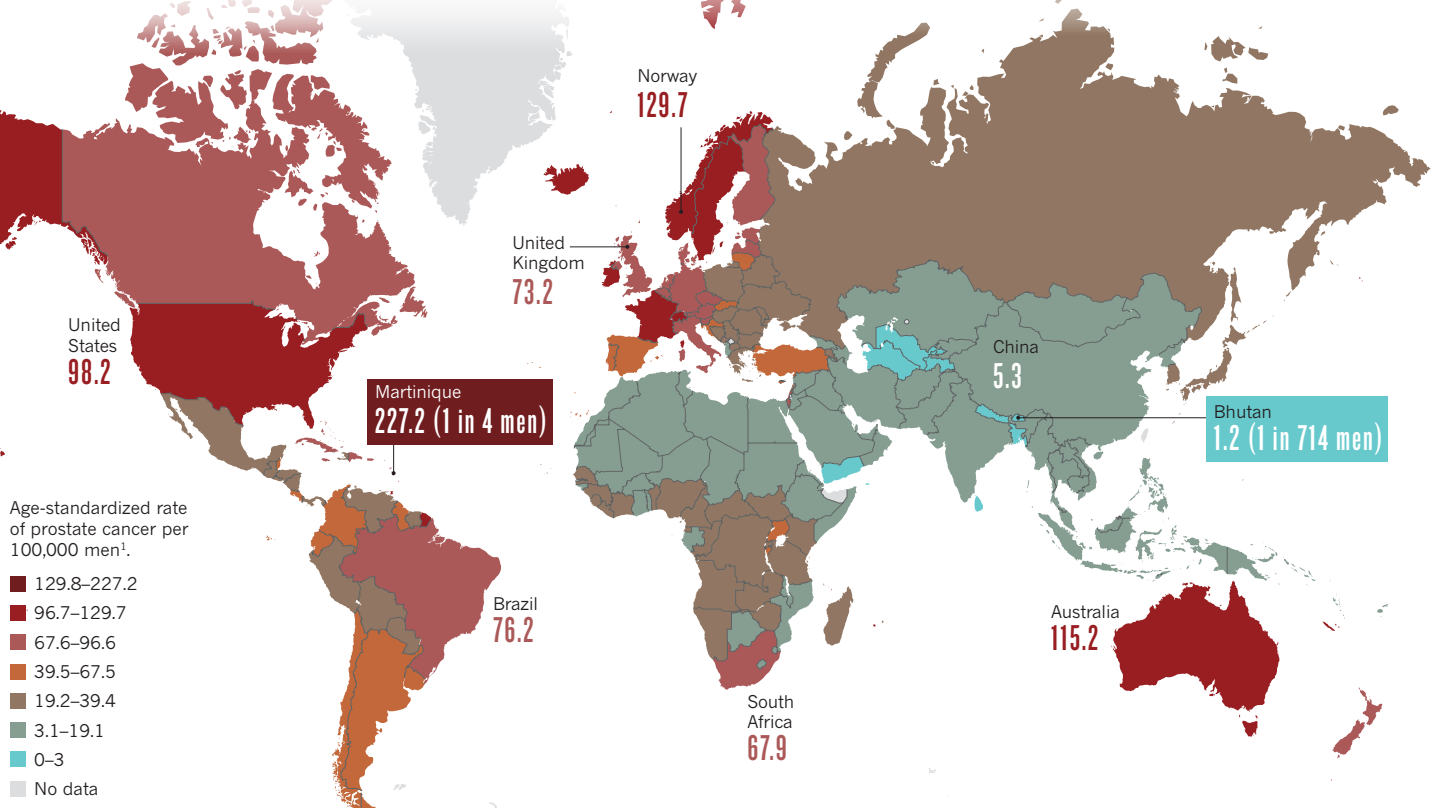
Feedback@nature.com  
Copyright © 2015 Nature Publishing Group

# SMALL ORGAN, BIG REACH

Prostate cancer is one of the most common cancers in men — most will develop the disease if they live long enough. But it is not always deadly, and the number of cases often depends on how hard doctors look for it. By **Richard Hodson**, infographic by **Mohamed Ashour**.

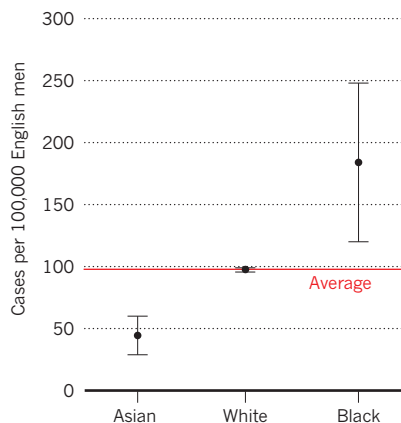
## GLOBAL INFLUENCE

The rate of prostate-cancer diagnosis varies more than 25-fold around the world. The incidence rate in a country is influenced by trends in diagnostic testing, which vary from place to place, as well as by the age and ethnic mix of a population.



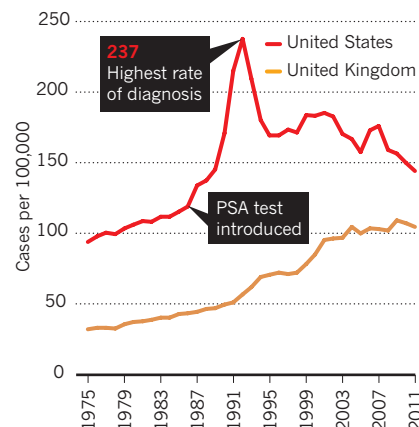
## ETHNICITY EFFECTS

On the Caribbean island of Martinique, men have a 26% chance of being diagnosed with prostate cancer by age 74 — the highest in the world. But in Bhutan, the risk is just 0.14%. Ethnicity may play a part. English black men have much higher rates of the disease than Asian men<sup>2</sup>.



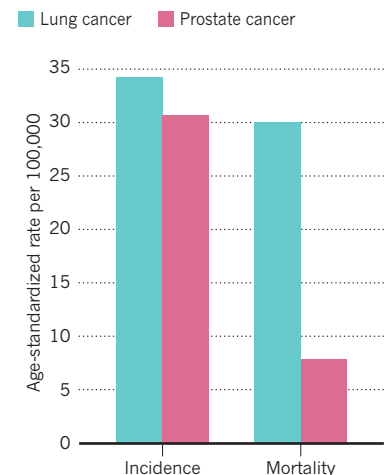
## LOOKING FOR TROUBLE

The rate of prostate-cancer diagnosis in the United States spiked after the prostate-specific antigen (PSA) test was introduced in 1986 (ref. 3). Testing men without symptoms is no longer recommended. In places where the test is used less, such as the United Kingdom, rates have increased only gradually<sup>4</sup>.



## HOW DEADLY?

Prostate cancer is the second most common cancer in men worldwide, just behind lung cancer. But for every 30 lives lost to lung cancer, just 8 men will succumb to prostate cancer<sup>1</sup>.





## A MATTER OF TIME

Age is the greatest risk factor for prostate cancer. Most (97%) prostate cancers occur in men over 50. As they get older, men are more likely to develop prostate cancer<sup>5</sup>.



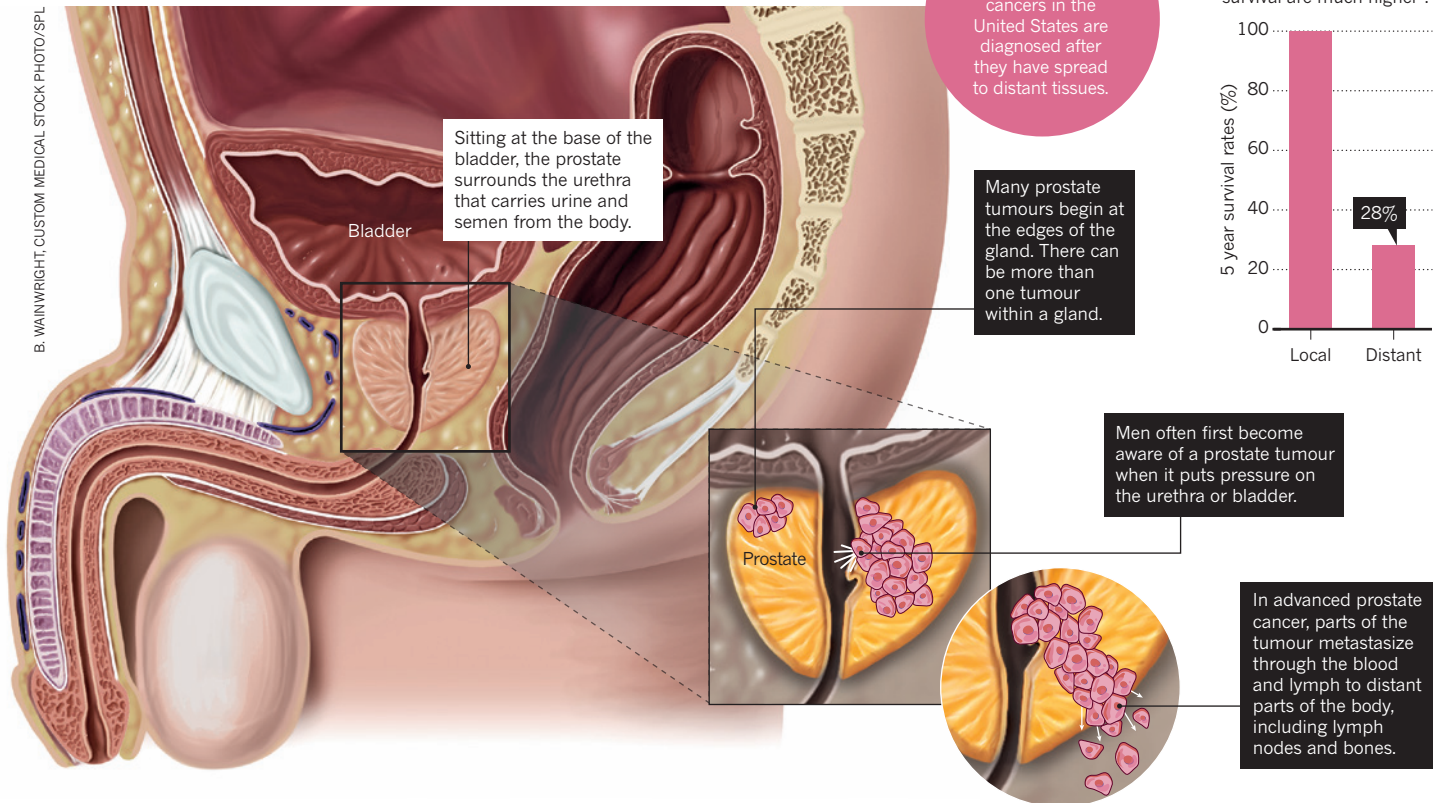
# 80s

For British men in their eighties, rates of prostate cancer fall. This may be a reflection of lower rates of PSA testing in this age group.

## MAN ON THE INSIDE

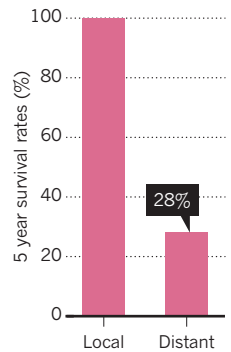
The prostate gland is a male organ involved in sexual function. Its size ranges from that of a walnut to that of a small apple, and can become enlarged as a result of cancer, inflammation or benign prostatic hyperplasia.

B. VAINWRIGHT, CUSTOM MEDICAL STOCK PHOTO/SPL



## SURVIVAL STORY

When prostate cancer is diagnosed early, before it has spread, chances of survival are much higher<sup>5</sup>.

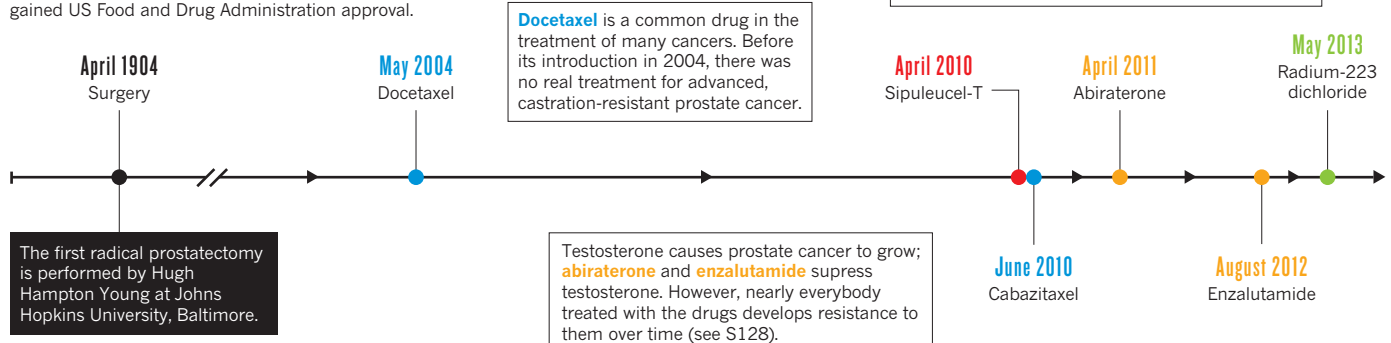


## A CENTURY OF TREATMENT

For localized prostate cancer, the most common intervention is surgical removal of the prostate — radical prostatectomy. If cancer has spread beyond the prostate it cannot be cured. Suppressing male hormones slows growth, but tumours can become resistant. Since 2004, therapies to target resistant metastatic cancer have gained US Food and Drug Administration approval.

- Chemotherapy
- Immunotherapy
- Hormone therapy
- Radiotherapy

The only approved immunotherapy for prostate cancer, **sipuleucel-T**, is costly and extends life by only a few months. Successors are in development, and combining them with other therapies may prove fruitful (see S134).



Sources: 1. International Agency for Research on Cancer; 2. National Cancer Intelligence Network; 3. National Cancer Institute's Surveillance, Epidemiology, and End Results Program; 4. Cancer Research UK; 5. Siegel, R. L. et al. *CA Cancer J. Clin.* **65**, 5–29 (2015).



A simple blood test is used to measure prostate-specific antigen, or PSA, but researchers continue to debate how best to use the test to save lives.

## SCREENING

# Diagnostic dilemma

*The standard blood test for prostate cancer led to a spike in diagnoses of the disease. But the technique's results are often misleading — and conflicting studies have not helped to forge a consensus.*

BY EMILY SOHN

**I**t was an appealing idea: a simple blood test that could detect prostate cancer early, before it could become life threatening. So appealing, in fact, that enthusiasm for the prostate-specific antigen (PSA) test caught on long before there was strong evidence to support it.

PSA is a protein that is produced by the prostate gland and is usually found in the blood at higher levels when a prostate tumour is present. The US Food and Drug Administration initially approved the PSA test for cancer monitoring in 1986, and by 1992 the US incidence rate for prostate cancer had more than doubled from 119 to 237 cases per 100,000 people. From 1992 to 2012, deaths from prostate cancer halved, from about 39 cases per 100,000 people to 20. “When you look at the curves, there’s nothing else like it with other cancers,” says Laurence Klotz, a urologic oncologist at the Sunnybrook Research Institute at the University of Toronto in Canada.

But scepticism also emerged early and deepened over time, especially as two closely watched trials produced drastically different results — one showing a substantial benefit of

screening and the other showing no benefit at all. In the meantime, studies have shown that whereas many men have had their lives saved by early detection with the test, many others have been diagnosed and treated for cancers that in all likelihood would never have caused them harm.

“The PSA was a genie that got out of the bottle well before randomized trials were initiated,” says Michael Barry, a primary care doctor at Massachusetts General Hospital in Boston and president of the Informed Medical Decisions Foundation, a Boston-based organization that advocates for evidence-based shared decision-making between doctors and patients. Even when trial results became available, he adds, they failed to resolve the question of whether the test was worthwhile.

Hundreds of studies have now analysed the consequences of screening. The results have led to important developments in testing protocols, treatment decisions and public trust in the PSA test. And taken together, these findings are starting to reveal how best to use the test to help more people and harm fewer of them. Still, researchers and clinicians continue to debate everything from what level of PSA

should be considered alarming to which men should have the test in the first place. Some 30 years after the PSA test was introduced, the question it raised still lacks a definitive answer — what is the best way to protect men from prostate cancer without treating those who are better off left alone?

“You can support any argument you want depending on which data you quote,” says Klotz. “We are not nearing consensus.”

## THE GOOD, THE BAD AND THE OPINIONS

PSA emerges from the prostate and circulates through the blood at levels that become increased for various reasons. By the mid-1980s, it was clear that prostate cancer was one of those reasons, and doctors began using the test to track progression of the disease. One of the first studies to suggest that the PSA test might also revolutionize the ability to screen for cancer emerged in 1991 (ref. 1), when researchers found that the test detected many more cancers than did rectal examination, which at the time was the best screening method available.

The study included more than 1,600 men who received the PSA test, which was



followed up with rectal exams and ultrasound scans if PSA levels were deemed high, as well as 300 men who underwent biopsies after being flagged during the course of clinical care. Of the 37 men in the study group who were diagnosed with prostate cancer, 12 of them would have been missed if they had received only rectal exams.

At the time, nearly 20% of men diagnosed with the disease had an advanced form that had already spread outside the prostate, so doctors were eager for a way to pinpoint the disease at an earlier, more treatable stage. The new findings offered hope that the PSA test might be the answer. When the study came out, media coverage was enthusiastic, and lead author William Catalona appeared on the television talk show *Good Morning America*. “I think that kind of kicked off the PSA era,” says Catalona, who is director of the clinical prostate-cancer programme at Northwestern University Feinberg School of Medicine in Chicago, Illinois.

Fritz Schröder, a professor of urology at Erasmus University Medical Center in Rotterdam, the Netherlands, remembers hearing of the study’s results with excitement and meeting with a colleague in Belgium to discuss them. Recognizing a clear need for a randomized trial to assess the PSA test’s ability to save lives, they put together the European Randomized

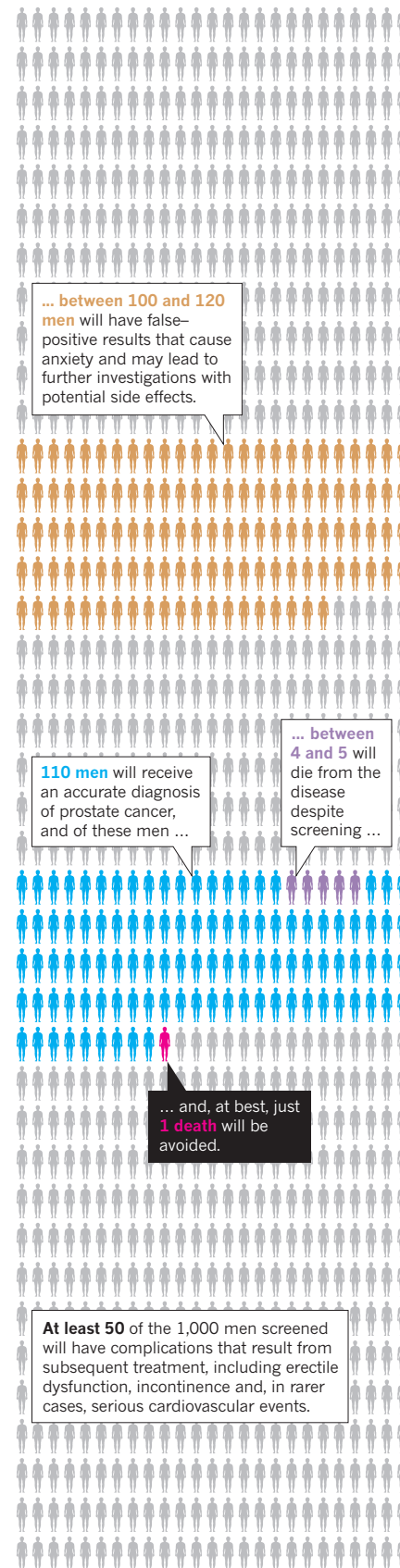
**“The PSA was a genie that got out of the bottle well before randomized trials were initiated.”**

Study of Screening for Prostate Cancer (ERSPC). This trial eventually grew to include 240,000 men from eight countries, who were randomly assigned into control and test groups, with the latter receiving PSA tests every one to four years. The first results from the ERSPC, which were published in 2009 and included nine years of follow-up data, reported a 20% drop in deaths from prostate cancer as a result of early detection with the PSA test<sup>2</sup>. In 2014, that figure grew to 27% after analyses were adjusted to include only men who had actually complied with the screening regimen to which they were assigned<sup>3</sup>.

Other lines of evidence have emerged to support screening. Since the beginning of widespread PSA testing, Catalona says, there has been an 80% drop in the percentage of patients in the United States whose cancers are metastatic at the time of diagnosis — a major factor in the declining US death rate from the disease. Trends are similar in other countries that have adopted screening, Catalona adds, with a link between when screening started and when death rates began to drop. Denmark, for example, started encouraging screening later than did other Nordic countries, and Danish prostate-cancer mortality levelled off later than in those neighbouring countries. Other researchers disagree about how many of those lives were saved as a result of the PSA

## TO SCREEN OR NOT TO SCREEN

Screening for prostate cancer with the prostate-specific antigen test produces an array of outcomes. If 1,000 men between the ages of 55 and 69 are screened every 1 to 4 years for a decade then ...



test because treatment has also improved during the same period.

Still, PSA-test advocates also point to an intangible benefit: peace of mind for men whose result indicates a low risk of prostate cancer. In two studies — one of men in their 40s and the other of men aged 60 — a PSA value of below 1 has been linked to a very low likelihood of developing aggressive cancer for many years afterwards. “There is no other biomarker,” Klotz says, “that gives you a 20-year predictive value of getting a common cancer.”

As encouraging as these findings may sound, consensus on them has been maddeningly elusive. When the ERSPC published its first results, the same journal issue published conflicting findings from another large trial, which included more than 76,000 US men who were randomly assigned to two groups: one that received a PSA test and rectal exam, and one that did not. This study, the US Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, showed no reduction in deaths from PSA testing after 11 years of follow-up<sup>4</sup>.

The two contradictory trials remain at the centre of debates about the benefits of PSA screening. The PLCO trial in particular has been criticized for widespread failure of subjects to comply with experimental conditions. Many men in the control group had a PSA test, whereas many assigned to the screening group went unscreened. Without adjustments to account for compliance rates, critics argue that the two groups were essentially the same. Still, the results continue to be included in major reviews, including the most recent analysis by the US Preventive Services Task Force (USPSTF), an independent panel of experts based in Rockville, Maryland, that makes evidence-based recommendations about preventive services.

The European trial has not escaped criticism, however. Even a 20% reduction in risk of death would add up to just one fewer case of metastatic prostate cancer per 1,000 men screened over 13 years. That short time horizon is problematic, says Barry, who argues that 13 years is not sufficient to assess the long-term effects of a test on a disease that often occurs later in life. Moreover, during the same time period that the PSA tests were being introduced, drastically improved treatments were entering the clinic — a development that may well account for better outcomes. “People accept that there is some benefit due to screening,” Barry says, “but how much is a subject of debate.”

## BETTER NOT TO KNOW?

Whatever their magnitude, the benefits of PSA screening come with some serious downsides. One complicating factor is that PSA levels can be increased for reasons that have nothing to do with cancer, including urinary-tract infections, inflammation and enlargement of the prostate, a benign

condition that becomes increasingly common as men age. As a result, many men who are flagged for follow-up by their high PSA levels do not have cancer at all. It is also common for men to harbour non-aggressive, slow-growing tumours for many years, and to eventually die from some other cause — which means that even men who do have cancer often do not need to know about it. The authors of one autopsy study found prostate cancers in 64% of men in their 60s who had died of something other than prostate cancer<sup>5</sup>. In the United States, a man has about a 14% risk of developing prostate cancer in his lifetime, but less than a 3% risk of dying from it. As a result, attempts to catch aggressive prostate cancers early have ensnared many men who never should have become cancer patients in the first place.

The discovery of an increased PSA level presents an array of potential risks. Biopsies can cause pain, fever, blood in the urine and infections, which are increasingly resistant to antibiotics. And because biopsies sample only a fraction of the prostate, they are not regarded as conclusive — uncertainty that often means further tests and biopsies even after a negative result. “Patients are happy to get a blood test,” says Peter Albertsen, a urologist at the University of Connecticut Health Center in Farmington, “but that starts the ball rolling downhill, and it can lead to all sorts of consequences.”

Wherever screening is widespread — including the United States, Australia and parts of Europe — unnecessary treatments are rampant, says Barry. The US National Cancer Institute estimates that for every 1,000 men screened regularly with the PSA test over the course of a decade, as many as 120 will get a false-positive result that may lead to a biopsy. Another 110 will get a cancer diagnosis (see ‘To screen or not to screen’). And nearly half of those 110 will have complications from treatment, including incontinence and sexual dysfunction. “Overdiagnosis,” Schröder says, “occurs at a rate that we find very disturbing.”

Cancer diagnoses carry an understudied psychological burden, Barry adds, and anxiety can linger even after reassuring biopsy results. There are hefty financial costs, too. In a 2011 analysis of data from the ERSPC that was extrapolated to the United States, researchers estimated that preventing one death from prostate cancer costs more than US\$5 million in screening, biopsies and treatments<sup>6</sup>. “If we treat patients for cancer and they die the same day they were destined to die from a heart attack,” says co-author Alex Shteynshlyuger, a urologist in private practice in New York, “what good have we achieved?”

Based on the seemingly high rate of potential

harm, the USPSTF updated its recommendations in 2012 to advise against routine PSA screening for all men. The United Kingdom has also decided against a national prostate-cancer screening programme owing to a lack of convincing evidence to support the PSA test. Still, other doctors and organizations continue to recommend screening, with variations in what age it should begin, how frequently tests should occur and what PSA levels should be considered concerning. The result is confusion for men who want to make informed decisions about their health.

## BLAMING THE MESSENGER

As scientists grapple with the data, there is another ongoing problem: the data keep changing because doctors are getting better at selecting the most eligible patients for screening and treatment. People are also making different choices about screening, with drops in both the number of US men having the PSA test and the number of prostate-cancer diagnoses, according to two studies published in November<sup>7,8</sup>. Still, disagreement persists about where the balance lies, and those arguments continue to rely on trial data that are becoming obsolete. “There are very few things we are doing today that we were doing the same way when the studies began,” Shteynshlyuger says. “The tectonic plates keep moving under our feet.”

Part of the shift is a result of advances in screening, which are helping doctors to zero in on aggressive cancers that need the most attention. Among the new strategies is a tool called the prostate health index (PHI), which measures three types of PSA. According to some research, the PHI is three times more specific than the standard PSA test, an improvement that reduces the number of unnecessary biopsies. Doctors around the world also now factor in a tumour’s Gleason score, which assesses aggressiveness based on the way that cancer cells look under a microscope. And researchers are continually re-examining the level at which the quantity of PSA in the blood should be considered abnormal. Some evidence, for example, supports the idea that the threshold for concern should be raised from its present value of 3–4 nanograms per millilitre to 10 nanograms per millilitre. Beyond PSA, scientists are also using magnetic resonance imaging to guide biopsies making false negatives less likely, as well as genetic tissue tests to screen for biomarkers that signal a cancer’s degree of aggressiveness (see page S124). These tests can be expensive, and health-insurance companies in the United States do not necessarily cover them. Many are so new, Barry adds, that there are insufficient data on outcomes. Rushing to accept newer tests before sound trial evidence arrives, in other words, might bring a repeat of the troubled PSA era all over again.

But the real crux of the screening debate is

what happens when results come in — and that is where big changes are happening. Within the past decade, for example, there has been a major spike in the number of men with low-risk cancers who choose to forgo treatment, instead taking a wait-and-see approach known as active surveillance, which, depending on the situation, could mean periodic screening or careful observations of symptoms (see page S126).

In 2006, 90% of US men diagnosed with prostate cancer were treated for it, says Stacy Loeb, a urologist at New York University School of Medicine. Today, only 50–60% opt for treatment. Sweden has been particularly quick to adopt the strategy: 91% of men with very low-risk and 74% of men with low-risk prostate cancer in the country now opt for active surveillance. As fewer men are treated, one hope is that the benefits of PSA testing will begin to outweigh the harms. “There is a lot of controversy about screening because it used to be done in such a very rudimentary fashion,” Loeb says. “We have come to recognize that it’s not so black and white.”

Given the uncertainties, many experts now recommend an approach that considers each patient’s situation individually. Statistical tools are helping with the process; at the University of Texas in San Antonio, for example, researchers used data from thousands of biopsies to create an online calculator that incorporates age, race, family history, PSA score and other factors into a recommendation that doctors and patients can consider together. This kind of shared-decision-making strategy is currently recommended by organizations such as the American Urological Association.

Forthcoming data may soon make screening decisions even more informed. In January 2016, researchers are expected to release 10-year follow-up results from the Prostate Testing for Cancer and Treatment (PROTECT) trial, which includes more than 1,600 British men who were diagnosed with localized prostate cancer using PSA tests and then randomly assigned to one of three treatment options, including active surveillance. But based on the history of PSA testing, it is hard to imagine that any fresh results will settle disagreement about screening once and for all. ■

*Emily Sohn is a freelance journalist in Minneapolis, Minnesota.*

1. Catalona, W. L. *et al.* *N. Engl. J. Med.* **324**, 1156–1161 (1991).
2. Schröder, F. H. *et al.* *N. Engl. J. Med.* **360**, 1320–1328 (2009).
3. Schröder, F. H. *et al.* *Lancet* **384**, 2027–2035 (2014).
4. Andriole, G. L. *et al.* *N. Engl. J. Med.* **360**, 1310–1319 (2009).
5. Sakr, W. A. *et al.* *In Vivo* **8**, 439–443 (1994).
6. Shteynshlyuger, A. & Andriole, G. L. *J. Urol.* **185**, 828–832 (2011).
7. Jemal, A. *J. Am. Med. Assoc.* **314**, 2054–2061 (2015).
8. Sammon, J. D. *J. Am. Med. Assoc.* **314**, 2077–2079 (2015).



## PERSPECTIVE



# Enforce the clinical guidelines

Prostate-specific antigen is not a bad test, it is just improperly applied, says **Monique Roobol**.

MARCO DE SWART

In developed countries, prostate cancer is the most common cancer in men (excluding non-melanoma skin cancer). In the United States alone, there will be 220,800 new cases and about 27,540 deaths from the disease in 2015 (ref. 1).

Not all prostate cancers are the same. Some cases are very aggressive, causing painful bone metastases and turning deadly, whereas others can stay dormant throughout the patient's life. This means that prostate cancer is only the second biggest cause of cancer deaths in US men, behind less-common lung cancer. So although a man's lifetime risk of being diagnosed with prostate cancer is 1 in 7, the risk of dying from prostate cancer is only 1 in 38.

A lot of prostate cancers are, therefore, overdiagnosed: they are unlikely to ever cause harm, let alone death. This overdiagnosis is initiated by the liberal application of a cheap, easy to apply and sensitive blood test: the prostate-specific antigen (PSA) test. And, crucially, that this test is given to too many men or too often, against best-practice guidelines.

To understand the current situation, it is helpful to outline the history of the test. From the mid-1980s until the early 1990s, PSA was officially used only to monitor the course of prostate cancer in men who were already diagnosed. At the time, prostate cancer was a life-threatening disease: one in every two or three patients died. In 1994, a team from Washington University School of Medicine in St Louis, Missouri, showed that adding a PSA test to a digital rectal examination increased the rate of early detection of the cancer — when the disease is confined to the prostate — by 78% (ref. 2). The same year, the US Food and Drug Administration approved this test combination to help detect cancer, and it was rapidly adopted. Physicians were able to actively seek out the disease, and it soon became clear that prostate cancer was actually very common.

These findings raised two questions. First, is it possible to reduce prostate-cancer mortality if the PSA test is introduced as a screening tool? And second, is it possible to reduce the side-effects of PSA screening, including overdiagnosis? To address these questions, two randomized trials — one in the United States<sup>3</sup> and one in Europe<sup>4</sup> — were initiated. Both trials have reported on the effect of PSA testing on prostate-cancer mortality several times over the years, and have always contradicted each other (although it is generally accepted that within the US trial contamination substantially limited researchers' ability to identify a clinically significant screening benefit). This lack of consensus and the considerable risk of overdiagnosis associated with PSA-based screening are the main reasons that screening for prostate cancer is still highly controversial, and why there are so few population-based government-initiated screening programmes.

What has become much clearer, however, is how to use the PSA test in such a way that the side effects are reduced. There are numerous papers describing how and when to use the PSA test. One of these outlined five golden rules<sup>5</sup>. PSA testing should not be carried out without pretest

counselling and explicit consent. Do not test in circumstances where screening clearly has no benefit — if a man has an estimated life expectancy of less than 10–15 years, or if he is over 60 years old and has a PSA-level of less than 1 nanogram per millilitre. The decision to perform a prostate biopsy — the next stage in a cancer diagnosis — should be taken based on multiple parameters and not solely on the PSA level. And a diagnosis of prostate cancer should not automatically lead to treatment.

Most of these recommendations have been included in the various national or regional guidelines on prostate-cancer screening, but are not being followed. American Urological Association (AUA) guidelines published in August stated that “screening patterns have been inappropriate and require modification”<sup>6</sup>. The same holds for Europe, where modern screening practices go against the European Association of Urology (EAU) guidelines. Notably, the highest screening rates are seen in men aged 75 or older, and men with a PSA of less than 1 nanogram per millilitre are being tested much too frequently<sup>7</sup>.

There are benefits to using the PSA test, including a reduction in incidence of metastatic disease<sup>8</sup> and in prostate-cancer mortality. But too many physicians are applying the test opportunistically and inappropriately. Doing so only highlights the much-debated drawbacks. But, when used judiciously and according to a fixed algorithm, these flaws can be avoided.

The time has come to actually implement the evidence-based guidelines into clinical practice. Medical associations should better communicate the best practice around PSA testing and strengthen the education of doctors — particularly general practitioners (GPs) who are usually the first point of contact, but are rarely up to date with the latest publications. GP requests for testing should be actively monitored to ensure the message is understood, rather than waiting for registry data to see if there has been an effect.

There is ample knowledge of how to streamline individual testing of men who have been appropriately informed. The PSA test is a key part of the urologist's toolkit. By implementing the EAU and AUA guidelines on prostate-cancer screening into clinical practice and stopping its misuse, we can prevent the loss of a screening test that has the potential to bring benefit to many men. ■

**Monique Roobol** is an epidemiologist at the Department of Urology, University of Erasmus, Rotterdam, Netherlands.  
e-mail: [m.roobol@erasmusmc.nl](mailto:m.roobol@erasmusmc.nl)

1. American Cancer Society. *What are the Key Statistics About Prostate Cancer?* [go.nature.com/lu3JwJ](http://go.nature.com/lu3JwJ) (ACS, 2015).
2. Catalona, W. J. *et al. J. Urol.* **151**, 1283–1290 (1994).
3. Andriole, G. L. *et al. J. Natl Cancer Inst.* **104**, 125–132 (2012).
4. Schröder, F. H. *et al. Lancet.* **384**, 2027–2035 (2014).
5. Vickers, A., Carlsson, S., Laudone, V. & Lilja, H. *Eur. Urol.* **66**, 188–190 (2014).
6. Eggener, S. E., Cifu, A. S. & Nabhan, C. J. *Am. Med. Assoc.* **314**, 825–826 (2015).
7. Nordström, T. *et al. Eur. Urol.* **63**, 419–425 (2013).
8. Schröder, F. H. *et al. Eur. Urol.* **62**, 745–752 (2012).

THE TIME HAS  
COME TO ACTUALLY  
IMPLEMENT  
THE EVIDENCE-BASED  
GUIDELINES INTO  
CLINICAL  
PRACTICE.



issue with prostate cancer is not necessarily detecting it early enough, but predicting which cancers are aggressive and which are indolent,” says Vadim Backman, a biomedical engineer at Northwestern University in Evanston, Illinois.

This conundrum has led to a difference of opinion about how widely to use the PSA test, and what action to take if cancer is detected (see page S120). PSA screening had quickly become widespread in the United States, but in 2012 the US Preventive Services Task Force recommended against routine screening.

The controversy has, however, also stimulated scientific creativity. Researchers are improving the way in which men who are likely to have aggressive forms of prostate cancer are identified to reduce unnecessary biopsies. And to cut down on needless treatment, they are developing better ways to evaluate biopsy tissue and determine which tumours truly pose a threat. Because prostate cancer has a long natural history, definitive studies to address these issues take a long time to complete. But a flurry of publications over the past few years, as well as the commercial introduction of several tests, suggest that scientific patience is paying off.

#### BEYOND PSA

To do a better job of deciding which men should have prostate biopsies, physicians need non-invasive tests, either to supplement PSA screening or to replace it entirely. “The most pressing need is to identify biomarkers that are specific for high-grade cancer,” says urologist Scott Tomlins at the University of Michigan in Ann Arbor. An ideal biomarker would only be expressed in prostate tissue, not elsewhere in the body, and only be found in aggressive cancer, not low-grade disease. To be useful as a screening test, the biomarker would also need to show these patterns in blood or urine, not just intact tissues<sup>1</sup>.

One approach is to improve on the concept of the PSA test with tests that can be used to spot patterns in particular forms of PSA or suites of other, related molecules in the blood that are more specifically linked to aggressive prostate cancer. One version of this approach, the prostate health index, integrates three forms of PSA into a single score, which is then used to determine the risk of an aggressive tumour. Whereas another, the 4Kscore test, measures a panel of four molecules, including two forms of PSA that, like PSA measured in the established test, belong to a group of enzymes called kallikreins.

A study of biopsy tissue from more than 6,000 men found that screening using the four-kallikrein panel could reduce the number of unnecessary biopsies — 43% fewer biopsies compared with the standard PSA test and a delay in the diagnosis of only a handful of aggressive cancers<sup>2</sup>. Another study bolsters these results. Researchers followed a cohort of men for more than 15 years, and found that the blood test predicts which men are more likely to develop metastatic prostate cancer in the long

#### PROGNOSIS

# Proportionate response

*Work to determine which prostate cancers are truly dangerous may finally be coming to fruition.*

BY SARAH DEWEERDT

A little knowledge can be both a blessing and a curse. Ever since the prostate-specific antigen (PSA) test was introduced in the United States as a method of screening for prostate cancer in the mid-1990s, physicians, scientists and public health officials have been wrestling with the problem of how to use it.

The blood test looks for high levels of

PSA — an enzyme that thins the semen to allow sperm to swim freely — and enables early detection of one of the most common forms of cancer. But it is far from infallible. A higher than average PSA reading is not necessarily the work of a malignant tumour, so the test flags many men who do not have cancer. And because prostate cancer is often indolent, meaning it is slow-growing and unlikely to spread, many of the cancers that are detected would never have threatened a man's health if left untreated. “The



term<sup>3</sup>. “I think that sends a very strong message that the way we measure this actually predicts something biological and disease-relevant,” says Hans Lilja, a clinical chemist at Memorial Sloan Kettering Cancer Center in New York and a leader of both studies.

Scientists are also developing urine-based screening tests. One of these tests measures levels of the biomarkers TMPRSS2-ERG and PCA3. About 80% of men with prostate cancer have at least one tumour that produces TMPRSS2-ERG — the result of a genetic scrambling that occurs very early in the development of many prostate cancers, leading to the fusion of two genes. PCA3, a molecule normally produced by prostate tissue, occurs at abnormally high levels in at least 90% of prostate cancers — but, unlike PSA, it is almost never elevated in benign conditions. “The unique thing about those is they’re very prostate-cancer specific,” says Tomlins.

Individuals with high levels of these two biomarkers in their urine tend to have a large amount of tumour in their prostate, Tomlins says. This itself is a good indicator that an aggressive cancer is at work. In May, his team reported that the two markers do a better job of zeroing in on aggressive cancers than the standard PSA blood test<sup>4</sup>. Next, they plan to test whether adding an assay for another molecule associated with aggressive cancer, SchlAP1, will further improve the test.

## MAKING SENSE OF SCREENING

More informative screening methods are good as far as they go, but researchers are also searching for another piece of the puzzle: how to improve the analysis of tissue taken in biopsies after a positive screening test. These advances would allow physicians to better distinguish which cancers require immediate treatment, and which can be monitored — an approach known as active surveillance (see page S126).

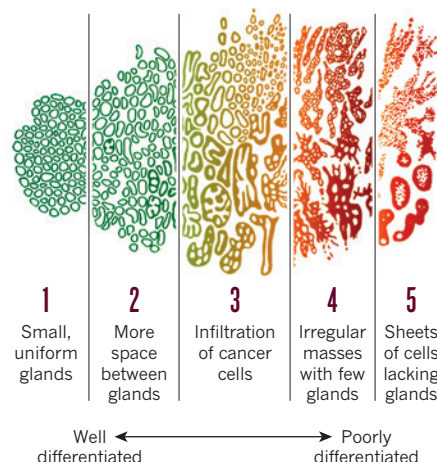
Oncologists currently evaluate prostate biopsies by Gleason grading, a method of scoring prostate tissue on a scale of 1 to 10 by how abnormal its cells appear (see ‘Scoring cancer’). Prostate tissue with a Gleason score of 5 or below is generally quiescent; tissue with a score of 8 or above requires immediate treatment.

But most common prostate tumours score a middle-of-the-road 6 or 7, and these are more vexing to deal with. Usually, grade 6 tumours can be safely managed with active surveillance. But a few will prove to be aggressive. Grade 7 tumours are more evenly split between those that are aggressive and those that are not. “There’s a lot in the kind of grey zone that we don’t know,” says Jack Cuzick, an epidemiologist at Queen Mary University of London.

Cuzick and his team have evaluated a biomarker known as the cell-cycle progression score, which measures the activity of genes related to cell division in biopsy tissue. The greater the rate of cell division, the more aggressive the tumour — a pattern that applies to

## SCORING CANCER

Prostate-cancer severity can be gauged by assessing how well differentiated the tissue appears under the microscope and grading it 1–5. This is done twice and the grade combined, giving a score of between 2 to 10.



many forms of cancer. In a study of 585 men with prostate cancer, the researchers showed that this approach provides additional information about which men with those intermediate Gleason-score biopsies are at risk of dying from prostate cancer over the course of ten years<sup>5</sup>. “Our feeling is that the cell-cycle progression score is a huge step forward to resolve many of the controversial cases,” says Cuzick, who also consults for Myriad Genetics, the maker of one cell-cycle progression test.

The cell-cycle progression score is one of an increasing number of genomic tests for prostate cancer — others evaluate between one- and two-dozen genes associated with prostate cancer. A weakness of these tests, however, is that they work best if they are applied to biopsies taken from the most aggressive part of the cancer — and that is not always obvious.

That is because many men with prostate cancer have multiple tumours of independent origin. These various tumours can differ in their aggressiveness. Studies of men who had their prostate removed have found that 15–40% of those diagnosed with low-grade cancer at biopsy actually have a more aggressive tumour elsewhere in the prostate.

Backman and his team have developed a form of microscopy that they say could overcome this difficulty by allowing pathologists to see changes inside cells that are too small to resolve with standard microscopy. The researchers, who formed NanoCytomics in Evanston, Illinois, to commercialize the technology, found that non-cancerous tissue taken from prostates that contain Gleason grade 6 tumours that turned out to be aggressive show characteristic nanoscale changes, especially in the packaging of DNA in the cell nucleus<sup>6</sup>. Prostates that contain non-aggressive grade 6 tumours do not show these alterations.

The advantage of this type of test is that doctors could potentially determine the

aggressiveness of a tumour without having to biopsy the tumour itself. “We don’t need to find the needle,” Backman says. “All we have to do is sample the haystack.”

## SMARTER BIOPSIES

Finely locating the tumours within the prostate is still an option, though. This is the focus of a third category of efforts aimed at improving the prostate-biopsy procedure, which involves taking samples of tissue — usually 10 to 12, but sometimes as many as 50 — with a fine needle.

Prostate biopsies have generally been performed with little information about exactly where in the prostate a sample comes from. This is because it is difficult to get a clear picture of the organ using standard imaging methods. But now, an approach known as multiparametric magnetic resonance imaging (MRI) is beginning to change that<sup>7</sup>. The procedure combines three techniques to generate a fuller picture of prostate anatomy and function. “We can go after the area that we think is most likely to have high-grade cancer,” Tomlins says.

Earlier this year, researchers found that oncologists locate more high-grade tumours when aided by multiparametric MRI than with standard biopsy procedures<sup>8</sup>. “It decreases the risks associated with active surveillance,” says the study leader Peter Black, a urological oncologist at Vancouver General Hospital in Canada. “You’re able to take these patients out of the active surveillance pool and treat them.”

The technique also makes it possible to follow the development of a specific tumour and repeatedly biopsy it over time. This capability could help to address some basic questions about prostate cancer — with implications for treatment. “For example, do low-grade tumours routinely turn into higher-grade or more aggressive tumours?” Tomlins says. “It’s a crucial question because it totally changes how we predict whether cancers are going to be indolent or aggressive.”

The challenge now is to bring together these varied strands of research, because the new biomarkers and testing strategies have largely been developed in isolation from each other. “Very little has been done to see if these can add to each other and how much we would gain by doing that,” Lilja says. So even as techniques that may yield a better understanding of a patient’s prognosis begin to roll out, scientists are aiming at the next round of improvements. ■

**Sarah DeWeerd** is a freelance science writer in Seattle, Washington.

1. Prensner, J. R., Rubin, M. A., Wei, J. T. & Chinnaiyan, A. M. *Sci. Transl. Med.* **4**, 127rv3 (2012).
2. Bryant, R. J. et al. *J. Natl Cancer Inst.* **107**, djv095 (2015).
3. Stattin, P. et al. *Eur. Urol.* **68**, 207–213 (2015).
4. Tomlins, S. A. et al. *Eur. Urol.* <http://doi.org/9hr> (2015).
5. Cuzick, J. et al. *Br. J. Cancer* **113**, 382–389 (2015).
6. Roy, H. K. et al. *PLoS ONE* **10**, e0115999 (2015).
7. Weinreb, J. C. et al. *Eur. Urol.* <http://doi.org/9hq> (2015).
8. Hamidreza, A. et al. *Urology* **85**, 423–429 (2015).



Having opted for active surveillance, Bill Wilson has avoided surgery and continues with his busy lifestyle.

never be; therefore, I did not have to immediately have aggressive treatment.”

Aggressive treatment is common: according to a recent report<sup>1</sup>, 50% of US men diagnosed with low-risk prostate cancer between 2010 and 2013 underwent radical prostatectomy (surgical removal of the prostate and surrounding lymph nodes). However, active surveillance is being offered by a growing number of physicians, and taken up by a rising number of men (see ‘Active prime time’). A large study in Sweden, led by Stacy Loeb, a New York University urologist, found that nearly half of men diagnosed with low-risk disease between 1998 and 2011 opted for active surveillance, with the proportion increasing over this time period<sup>2</sup>. In the United States, the turn toward active surveillance for low-risk disease has been dramatic, growing from single-digit percentages in the late 1990s to 40% between 2010 and 2013 (ref. 1).

Disagreements remain on which cancers are best suited to active surveillance, and on how to monitor those men selected. Nonetheless, active surveillance “has spiked and become prime time”, says Loeb. And for men whose disease does progress, new treatments and protocols are lengthening and improving the quality of their remaining life.

### A BALANCING ACT

According to criteria set by Carter’s Johns Hopkins group, Wilson was a good candidate for active surveillance: his tumour cells did not look aggressive under the microscope; the tumour could not be felt by digital rectal exam and was only picked up by needle biopsy; and he had relatively low levels of prostate-specific antigen (PSA), a blood marker that is a proxy for the presence of prostate cancer (Wilson underwent the biopsy only because his PSA was found to be slightly elevated: 0.57 nanograms per millilitre above the threshold of 4 ng ml<sup>-1</sup>).

The active-surveillance regimen requires Wilson to visit Johns Hopkins every six months for a digital rectal exam and a blood test. Annually, he undergoes either a biopsy or a magnetic resonance imaging scan, for a more detailed inspection. His most recent biopsy worried Wilson slightly because small areas of cancer were found in 3 of the 12 tissue samples taken, up from 2 of the 12 at diagnosis. Still, the cancer cells did not appear more aggressive, and his PSA levels remained reassuringly low.

The surge in uptake of active surveillance is due, in part, to the response of a medical community that was roundly criticized for overtreatment of a disease that is too readily identified by PSA screening (see page S120). By offering active surveillance as a conservative way to manage men at lower risk, clinicians hope that the balance of harms and benefits from PSA screening will shift. “The acceptance of surveillance is going to be a crucial piece of rehabilitating PSA screening,”

### TREATMENT

# When less is more

*Surveillance is becoming a watchword for men with less-aggressive prostate cancer. If and when the disease progresses, new and newly-timed therapies are at hand.*

BY MEREDITH WADMAN

When Bill Wilson learned that he had prostate cancer in 2011, he wanted to race to the nearest operating theatre. Wilson, a 71-year-old former IBM executive from St. Michael’s, Maryland, says his first thought was: “I’m going to get it out of there.”

But his urologist encouraged him to talk to other specialists, and so Wilson visited Balentine Carter, a prostate-cancer researcher

at Johns Hopkins University in Baltimore, Maryland. Carter urged him to consider a more conservative approach called active surveillance. This involved Carter simply monitoring the tumour over time; treatment would be launched only if the disease progressed.

“A huge weight was lifted off my shoulders immediately,” says Wilson, who had feared both the ordeal of surgery and its common side effects, including incontinence and impotence. “My cancer was not life-threatening and might



SOURCE: REF. 1

says Laurence Klotz, a urologist in Sunnybrook Health Sciences Centre at the University of Toronto, Canada.

Klotz's group has published some of the longest-term data on the results of active surveillance, which it began offering in 1995 (ref. 3). The Toronto group followed 993 men, most of whom had low-risk tumours. It also included some men with slightly higher-risk tumours who had other significant illnesses and less than ten years of life expectancy. The group was followed for a median of 6.4 years; 1.5% died of prostate cancer.

For men with low-risk tumours, the comparable outcome after surgical removal of the prostate is slightly better. A 2011 study involving 24,000 men showed that, after 15 years of follow-up, between 0.2% and 1.2% on average had died of prostate cancer<sup>4</sup>. But in the same study, men with slightly higher-risk tumours — equivalent to the riskier subgroup in the Toronto cohort — did not fare as well: their average 15-year prostate-cancer-specific mortality ranged from 4.2% to 6.5%.

Carter's team published its results in August<sup>5</sup>. Of the 1,298 men with low- or very-low-risk disease placed on active surveillance since 1995, only 0.1% had died of prostate cancer at 10 years of follow-up; at 15 years, the figure was unchanged.

"The take-home message is, in well-chosen patients, active surveillance is safe," says Fred Saad, a urologist at the University of Montreal Hospital Centre in Canada. However, the Johns Hopkins and Toronto studies also highlight one of the sticking points that active-surveillance advocates are still wrestling with: inclusion criteria. The Johns Hopkins group was more conservative than the Toronto team: in Baltimore, only low-risk men were enrolled, which could help to explain the better results. In a 2014 review of ten active-surveillance studies, the approach seemed to reduce over-treatment without compromising men's ten-year cancer survival, but the authors

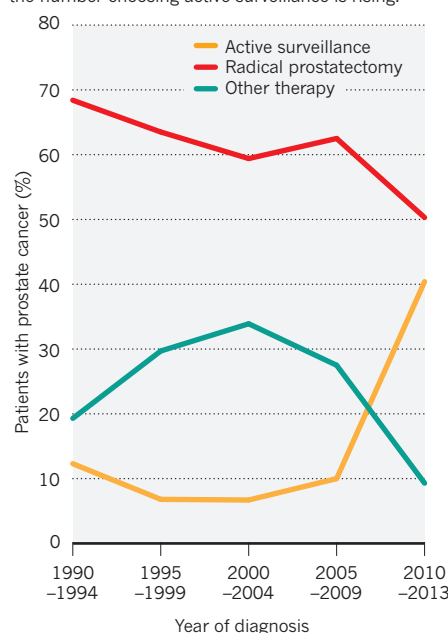
stated that the data are not yet mature enough for definitive conclusions to be drawn. Furthermore, they wrote bluntly, current tools for selecting patients and monitoring the disease are "inadequate and imprecise"<sup>6</sup>.

Freddie Hamdy, who specializes in prostate and bladder cancer at the University of Oxford, UK, notes that there are multiple active surveillance protocols in use at different centres, and many ways of interpreting them. "How are you going to detect or decide that the patient should no longer be on active surveillance because it's not safe?" he says. "That's the real challenge."

Hamdy hopes to shed new light on this question with a first-of-its-kind, randomized

## ACTIVE PRIME TIME

In US men diagnosed with low-risk prostate cancer, the number choosing active surveillance is rising.



controlled trial in which 1,643 men newly diagnosed with localized prostate cancer are randomly assigned to receive active monitoring, surgery or radiation therapy. Hamdy's group plans to report results of the ProtecT trial, showing disease-specific survival rates at a median ten years' follow-up, as early as spring 2016.

## SECOND ACTS

For men whose disease at diagnosis is too aggressive for active surveillance, treatment begins with some combination of surgery, radiation and medication — the first-line treatments for decades. Because prostate-cancer cells are stimulated to grow by testosterone, more than 90% of which is made in the testes, one approach is surgical castration: the operative removal of the testes. (The rest of the body's testosterone is made in the adrenal glands.) The same end can be achieved with medical castration — the use of drugs that suppress the release of hormones that stimulate testosterone production.

Nearly all tumours eventually become resistant to these testosterone-lowering approaches, and when they spread, they are described as metastatic castration-resistant prostate cancer. For men in this stage, the past five years have seen remarkable advances in both new, life-prolonging drugs and in better ways to use older agents.

The drugs abiraterone and enzalutamide were first approved in 2011 and 2012, respectively. Abiraterone works primarily by interfering with testosterone synthesis, whereas enzalutamide prevents the hormone from binding to the androgen receptor. Although neither is curative, the advent of these oral

agents has improved both survival and quality of life for men with advanced disease. "It is fantastic to give drugs that have almost no side effects and that make men feel better, have less pain and live longer," says Saad. Nonetheless, over time, most patients develop resistance to these drugs — a reality that has galvanized the hunt for new therapeutics (see page S128).

In another positive development, striking new evidence shows that a change in the timing of chemotherapy can buy men with metastatic disease many more months of life. Docetaxel — a chemotherapy agent that suppresses cancer-cell division — was approved in 2004 for use after medical or surgical castration has failed. But results of the CHARTED trial published this year showed that the lives of men with metastatic disease were significantly extended when they were given docetaxel earlier, simultaneously with castration<sup>7</sup>. They lived a median of 14 months longer than men who underwent only medical or surgical castration. For men with the most extensive disease, the difference was still more pronounced: 17 months. The results "were really dramatic," says Matthew Cooperberg, a prostate-cancer specialist at the University of California, San Francisco. "You don't see a 17-month survival advantage very often in these types of trials."

The CHARTED trial was led by Christopher Sweeney, a physician at the Dana-Farber Cancer Institute in Boston, Massachusetts. He suggests that the success came from deploying drugs with different mechanisms of action, thus targeting both the testosterone-sensitive and testosterone-insensitive cancer cells simultaneously. "That's speculation, but something like that might be happening," says Sweeney. He and others are now running trials to see whether patients will benefit when the newer drugs, enzalutamide and abiraterone, are likewise deployed alongside castration earlier in the course of metastatic disease.

Meanwhile, having been spared aggressive treatment, Wilson is taking the 'active' part of his active surveillance seriously. He spends his time sailing his 35-foot sloop, *Adagio*, training his dog Jeter for agility contests, and getting out into the countryside — his disease far from his mind. "I just came back from hiking in the mountains in Arizona and put the 40-year-olds to shame." ■

**Meredith Wadman** is a freelance writer based in Washington, DC, and an editorial fellow at *New America*.

- Cooperberg, M. R. & Carroll, P. R. *J. Am. Med. Assoc.* **314**, 80-81 (2015).
- Loeb, S. et al. *J. Urology* **190**, 1742-1749 (2013).
- Klotz, L. et al. *J. Clin. Oncol.* **33**, 272-277 (2015).
- Eggener, S. E. et al. *J. Urology* **185**, 869-875 (2011).
- Tosoian, J. J. et al. *J. Clin. Oncol.* <http://dx.doi.org/10.1200/JCO.2015.62.5764> (2015).
- Thomsen, F. B. et al. *J. Surg. Oncol.* **109**, 830-835 (2014).
- Sweeney, C. J. et al. *N. Engl. J. Med.* **373**, 737-746 (2015).



Scans that use  $\gamma$ -rays show the spread of cancer (white) from the prostate to the bones.

#### METASTASIS

# Resistance fighters

*Strategies to destroy treatment-defying tumours in men with prostate cancer are beginning to make a difference.*

BY NEIL SAVAGE

When the patient entered a trial of an experimental prostate-cancer treatment, he was in bad shape. The disease had spread to at least ten different parts of his body, including his arm and leg bones, and his hip, spine and ribs. The tumours caused him so much discomfort that, despite heavy use of pain-relieving medication, he was unable to sit up. Chemotherapy had failed to halt the spread of the cancer. But now, nearly seven years after finishing the trial, the patient's tumours have disappeared, his pain has vanished and his blood levels of prostate-specific antigen (PSA; a protein biomarker used to monitor malignancy) give no indication of the disease.

"We always are cautious using the word 'cure,'" says Fred Saad, a prostate-cancer researcher at the University of Montreal in Canada, who ran the study. "There are

diseases we have cured in a very advanced stage, like lymphoma, like testicular cancer," he says. But despite individual successes, advanced prostate cancer is still considered to be incurable.

Many men with the disease have tumours that grow so slowly that they never cause a problem. Others can be cured by treating the tumour within the prostate gland. But in some, the cancer spreads to elsewhere in the body, usually to the bones. The first line of treatment for these men is to suppress the male sex hormones (androgens), such as testosterone, that stimulate prostate tumours to grow — a form of chemical castration. Within a year or two, however, tumours become resistant to this treatment.

Until the early 2000s, there were no available treatment options for castration-resistant prostate cancer (CRPC). Since 2010, a handful of therapeutic strategies for treating CRPC have emerged. But at best, they add a few months to patients' median survival

time. So researchers are working to understand the mechanisms by which prostate cancer is able to resist efforts to overcome it, and to develop approaches that can permanently defeat the disease.

Saad's study is one such attempt<sup>1</sup>. The phase II trial focused on men with metastatic CRPC whose condition had worsened despite undergoing chemotherapy with docetaxel, a drug from the taxane family. The researchers focused on clusterin, a protein that increases in concentration when cells are stressed and seems to protect the cells from damaging agents. Researchers suspect that clusterin helps various types of tumour to become resistant to drugs used in chemotherapy. By inhibiting clusterin with a drug known as custirsen, the team hoped to once again make CRPC tumours vulnerable to the effects of chemotherapy.

#### REMARKABLE RESPONSE

The results of the trial were encouraging. Men who received custirsen together with docetaxel and the immunosuppressant drug prednisone showed a reduction in both pain and PSA levels. Saad's patient with the impressive results, who was 62 when he started the trial, had seen his PSA level shoot up from 74 to 115 nanograms per millilitre in the 3 weeks before treatment (a PSA level below 4 ng ml<sup>-1</sup> is generally considered normal; a man who has had his prostate removed and is now cancer free should have a level of 0). Within 2 weeks of starting the trial, his PSA levels had dropped to around 70 ng ml<sup>-1</sup>, and after 24 weeks, they had plummeted to less than 0.03 ng ml<sup>-1</sup>. Seven years on, the patient's PSA level is undetectable. Although this particular case does not prove that custirsen can cure prostate cancer, Saad thinks that it is remarkable.

The larger story of custirsen — an example of a DNA-based 'antisense' drug that binds to RNA and switches a gene off — is less clear. A phase III trial that used custirsen alongside docetaxel and prednisone showed no statistically significant improvement in the survival of participants with advanced prostate cancer compared with those who received the same treatment, but without custirsen. The results of another phase III trial, which combines custirsen and prednisone with a different anticancer drug, cabazitaxel, are expected by early 2016.

Saad says that the key to finding effective treatments for advanced prostate cancer lies in identifying those men — like his star patient — who will respond to a given therapy, perhaps because of a particular mutation or variation in their tumour. That requires determining which molecular mechanisms help to confer resistance to drugs in certain people, and finding ways to test for them. Large studies that are unable to identify subgroups of patients who respond



to a therapy can lead researchers to dismiss drugs that would work well in the right individuals. “The ones that are actually responding are drowned in a sea of non-responders,” says Saad.

### SPLICE VARIANTS

The resistance of prostate cancer to chemical castration develops by several routes. One biomarker of a particular mechanism of resistance has already been found — a receptor protein that binds androgens within the cell. Two new anti-androgen drugs, enzalutamide and abiraterone, can extend the lives of men with metastatic prostate cancer by up to three years. Eventually, those drugs stop working in almost all men — but 20–40% of patients never respond at all<sup>2</sup>. The reason for this initial resistance is a variation in the messenger RNA sequence that is used as a template for building the androgen-receptor protein itself.

To make the receptor, the DNA of the androgen-receptor gene is first converted into a sequence of RNA that encodes all parts of the receptor protein. Any RNA that does not code for protein is cut out and the remaining pieces of RNA are joined or ‘spliced’ together to produce the receptor template. Occasionally, pieces of protein-coding RNA are also removed during splicing, which creates different versions — splice variants — of the receptor template. In one, androgen-receptor variant 7 (AR-V7), the receptor is missing its ligand-binding area, called the carboxyl terminal. This is what the androgen normally attaches to, but with no receptor mechanism to interfere with, the drugs are powerless. However, the area of the androgen receptor that triggers the cell to divide, found at the protein’s opposite end, still works. “It can cause the cancer cell to grow and divide even without testosterone being present,” says Emmanouel Antonarakis, an oncologist at Johns Hopkins Sidney Kimmel Comprehensive Cancer Center in Baltimore, Maryland, who helped to identify the variant.

Using a blood test, Antonarakis has compared men whose tumours contain AR-V7 with those whose tumours do not. Whereas men who tested negative for AR-V7 responded equally well to both anti-androgen drugs and chemotherapy with taxanes, those with AR-V7 did not respond to the anti-androgen drugs. But they did respond to chemotherapy with taxanes, which disrupt the microtubules that help cells to divide. Antonarakis’s finding is supported by a study from the Erasmus University Medical Center Rotterdam in the Netherlands, in which investigators showed that AR-V7 does not diminish the effect of the taxane cabazitaxel<sup>3</sup>. A study from University Hospital Ulm in Germany confirmed the link between the variant and androgen resistance<sup>4</sup>. If these

findings hold up, Antonarakis says that men with AR-V7 could skip the anti-androgen treatment and go straight for chemotherapy. Men who test negative can choose between the two.

Soon, there might also be more treatment options for men with AR-V7. The drug galeterone, for example, the subject of a phase III trial, works in three different ways. Like enzalutamide, it prevents androgens from binding to their receptors. And like abiraterone, it interferes with the production of testosterone. But galeterone also degrades the androgen receptor itself — an action

**“If the receptor can’t bind to DNA, it can’t switch on these genes to divide, multiply and spread.”**

that could prevent the cell from becoming resistant to the other two lines of attack. According to Antonarakis, galeterone is the first anti-androgen drug “that actually may be effective in men who have AR-V7”. So far, testing has shown that PSA levels dropped in men with CRPC who took galeterone during phase II trials. Initial results of a phase III trial, which focuses specifically on men with AR-V7, are expected by the end of 2016.

Essa Pharma of Vancouver, Canada, is taking a different approach to the problem of resistance with its drug EPI-506, currently being prepared for phase I/II testing. Although most anti-androgen drugs target the end of the androgen receptor to which androgens bind, Essa’s drug is the first to target the receptor’s opposite end, which can interact with the DNA of the cell. By blocking this part of the receptor, the drug could prevent it from doing its job — stopping the cancer in its tracks. “If it can’t bind to DNA, it can’t switch on these genes to divide, multiply and spread,” Antonarakis says.

### DNA REPAIR

Splice variants are not the only way that prostate cancer can become resistant to anti-androgen drugs. When hit with a therapy, the disease — like any other cancer — mutates and develops mechanisms to help it to survive and grow. And anti-androgen drugs such as enzalutamide and abiraterone can inadvertently switch on the cancer-promoting mechanisms that androgens normally suppress. “You activate a sort of replacement pathway,” says Timothy Thompson, an oncologist who is director of prostate-cancer research at the University of Texas MD Anderson Cancer Center in Houston.

Anti-androgen drugs actually “unrepress” oncogenes such as c-MYB, switching on pathways that help to promote the growth of cancer. In fact, drugs such as enzalutamide seem to stimulate mechanisms that repair

DNA damage<sup>5</sup> — not enough to create normal cells, but sufficient to allow cancer cells to multiply and spread.

Researchers are searching for specific steps in the c-MYB pathway that they could target with new or existing drugs. Of particular interest is a class of enzymes called poly(ADP-ribose) polymerases, known as PARPs, which play a part in repairing damaged DNA<sup>6</sup>. Drugs that inhibit PARPs might disrupt the repair process and make cells more vulnerable to other forms of chemotherapy. PARP inhibitors are already being tested for the treatment of patients with breast cancer who have mutations in the genes *BRCA1* and *BRCA2*, and in December 2014, olaparib became the first such drug to be approved by the US Food and Drug Administration for treating ovarian cancers with the same *BRCA* mutations.

In April 2015, researchers from the Institute of Cancer Research and the Royal Marsden NHS Foundation Trust in London presented the results of a phase II trial of olaparib for men with metastatic prostate cancer. Lead researcher Johann De Bono says that a handful of patients showed “spectacular responses” to treatment with olaparib — their tumours disappeared from imaging scans. Others saw their PSA level cut in half. And all of the seven trial participants who had mutations in the gene *BRCA2* responded to the drug in some way.

Such discoveries could open the door to multipronged approaches in the fight against a disease for which there was no effective therapy just over a decade ago. That could revolutionize the treatment of advanced prostate cancer, says Saad, by bringing approaches in line with those for other cancers. “Prostate cancer is still one of the few, or only, solid tumours treated with a mono-treatment approach,” he says. “Where we need to go in the future is combining therapies.”

Although it might be a long time before the lives of most men with advanced prostate cancer can be significantly prolonged, Antonarakis agrees that combining therapies that block androgen receptors and destroy resistance mechanisms will soon stop the disease from being 100% fatal. “In the next five to ten years,” he predicts, “we will be able to cure a small percentage of metastatic castration-resistant prostate cancer.” ■

**Neil Savage** is a freelance science and technology writer in Lowell, Massachusetts.

1. Muhammad, L. A. & Saad, F. *Expert Rev. Anticancer Ther.* **15**, 1049–1061 (2015).
2. Emmanouel, S. et al. *N. Engl. J. Med.* **371**, 1028–1038 (2014).
3. Onstenk, W. et al. *Eur. Urol.* **68**, 939–945 (2015).
4. Steinestel, J. et al. *Oncotarget* <http://dx.doi.org/10.18632/oncotarget.3925> (2015).
5. Thompson, T. C. & Li, L. *Oncotarget* **5**, 8816–8817 (2014).
6. Livraghi, L. & Garber, J. E. *BMC Med.* **13**, 188 (2015).

## MICROBIOLOGY

# Inflammatory evidence

*Inflammation is an underlying cause of many cancers — and prostate cancer might turn out to be one of their number.*

BY KIRSTEN WEIR

When Angelo De Marzo peers at cancerous prostate tissue through the lens of his microscope, he often sees a total mess.

There are the cancer cells, of course, as well as abnormal cells thought to be precursors to cancer. There are also pockets of a third cell type: shrunk, withered cells that — despite their ailing appearance — are dividing rapidly. And surrounding that sickly stew are areas in which inflammation has set in for the long run.

But that might not be by accident. Inflammation in the prostate gland is common, and it is even more common in men with prostate cancer. De Marzo — a pathologist and oncologist at the Johns Hopkins University School of Medicine in Baltimore, Maryland — is part of a growing group of researchers who suspect that inflammation could be both a symptom and a cause of the disease. If so, physicians might one day be able to treat or even prevent prostate cancer by turning down the volume of the body's immune response.

## DOUBLE-EDGED SWORD

The immune system is a fickle friend. It protects us against invading pathogens and attempts to snuff out precancerous cells before they run wild. Inflammation lies at the heart of the immune response. But in the rush to attack potential pathogens, inflammation can cause collateral damage. "It's a two-edged sword," De Marzo says.

In the past two decades, scientists have begun to determine precisely how inflammation over an extended period of time could lead to the development of tumours. The classic example is gastric cancer, which can be caused by persistent

inflammation that is triggered by the bacterium *Helicobacter pylori*. Inflammation is also implicated in cancers of the liver, bladder and colon. As many as one-fifth of all cancers might be attributable to inflammation, according to Scott Lucia, a pathologist at the University of Colorado's Anschutz Medical Campus in Aurora.

Results from animals and humans suggest that prostate cancer also belongs on that list. "There's no definitive smoking gun that inflammation causes prostate cancer," says De Marzo, but "there's a lot of evidence building".

One reason for the uncertainty is that most samples of prostate tissue that are available for researchers to study have been removed from patients because of a medical problem — usually, in a biopsy performed after the discovery of an elevated level of prostate-specific antigen (PSA) in the blood. PSA is produced by the prostate gland, and a high concentration of the protein indicates that a person could have prostate cancer. But chronic inflammation alone can also raise PSA levels. As a consequence, men with inflamed prostates might be more likely to undergo biopsies that detect small tumours that would otherwise have gone unnoticed. If so, the association between inflammation and prostate cancer could be just an illusion.

De Marzo, Lucia and colleagues found a way to avoid this 'ascertainment bias' by using data from a fortuitously designed clinical trial. Between 1993 and 2004, the Prostate Cancer Prevention Trial set out to determine whether the drug finasteride could prevent prostate cancer. All participants who did not have cause for biopsy during the course of the trial were required to undergo an end-of-study biopsy, even if their PSA levels were low. By examining samples of benign tissue taken from the prostates of 400 men who were given a placebo in

the trial, around half of whom had been diagnosed with prostate cancer, De Marzo and

**Acne-causing bacteria are linked to prostate-cancer mortality.**

Lucia's team discovered that inflammation was very common<sup>1</sup>. Indeed, 78% of the men who were free from cancer showed signs of inflammation. However, inflammation was still much more likely to be found in men with cancer, appearing in 86% of samples from men with the disease, and 88% of samples from men with the most aggressive, high-grade cancer. "There is a relationship between cancer and inflammation," says Lucia. "As the amount of inflammation goes up, the odds ratio of having cancer — and in particular, high-grade cancer — went up."

Although De Marzo and Lucia's study confirmed an association between inflammation and prostate cancer, it was unable to answer the question of which comes first. "With something as common as inflammation, you see these relationships, but you don't know if they're causative," says Lucia. "If we could remove inflammation, would we lower the risk of prostate cancer? We don't have a means of doing that right now."

## INFECTIONS AND DIET

If inflammation does contribute to the development of prostate cancer, it is logical to ask what might be the cause. Infection is the leading suspect, and has been for some time.

In the 1950s, researchers observed that prostate cancer was more common in uncircumcised men<sup>2</sup>. This finding led them to propose that prostate cancer might be triggered by sexually transmitted pathogens, which they reasoned were more likely to be present in uncircumcised men. The hypothesis has since been supported by a number of



population-based studies. In particular, the bacterial infections gonorrhoea and chlamydia have been linked to an increase in the risk of developing prostate cancer, as has infection with the protozoan *Trichomonas vaginalis*.

Such infections can now be treated quickly with antibiotics. But rodent models hint that a short-term infection can launch what becomes an extended, or chronic, inflammatory response. Karen Sfanos, a pathologist at Johns Hopkins University School of Medicine, found that after a rat or mouse is cleared of a bacterial infection of the prostate, inflammation in the gland can persist for the rest of the animal's life. "Even a single infection seems to set up some kind of chronic inflammatory event," she says.

Sexually transmitted bacteria and protozoa are not the only pathogens that make their way into the prostate, thanks to the gland's location in the body. "The urethra actually passes through the prostate," Sfanos says. "There could be a very rich flora that's poised right there, where the prostate sits, that could continually be a source of exposure to microorganisms."

Sfanos has shown that strains of the bacterium *Escherichia coli* that are associated with urinary tract infections can cause an inflammatory response in the prostate of rodents. And so can *Propionibacterium acnes*, the bacterium associated with the common skin condition acne, according to studies in men. The culturing of *P. acnes* from inflamed prostate tissue led to the finding that men with a history of severe acne had a significantly increased risk of death from prostate cancer<sup>3</sup>.

Although infection is likely to cause chronic inflammation of the prostate, another suspect is the food on your plate. Prostate cancer is much more common in the United States and Western Europe than in Asia. "Diet could be one of the factors that explain the differences in rates," says Elizabeth Platz, an epidemiologist at Johns Hopkins Bloomberg School of Public Health.

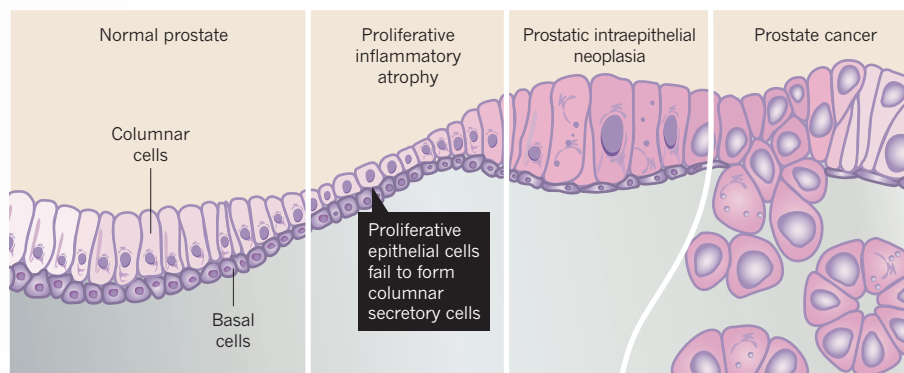
Research has shown that the consumption of certain foods can raise or lower the risk of developing prostate cancer. For example, a diet rich in red meat (and charred meat, in particular) seems to increase the risk. In a study by De Marzo and his team, rats that were fed PhIP — a carcinogenic compound that is abundant in well-cooked meat — developed cancer in the ventral lobe of the prostate<sup>4</sup>. Notably, the team also found that inflammatory cells were more plentiful in the same lobe. Foods with anti-inflammatory properties, such as soya beans and green tea, however, have been shown to decrease the incidence of prostate cancer in animals. Those foods have also been linked to a lower risk of developing prostate cancer in epidemiological studies in humans.

De Marzo thinks that a number of factors probably come together to create chronic inflammation in the prostate. "Something

**"Diet could be one of the factors that explain the differences in rates."**

## CANCER CULPRIT

Pockets of shrivelled cells called proliferative inflammatory atrophy may be a precursor to prostatic intraepithelial neoplasia and prostate cancer. These lesions are often associated with chronic inflammation.



seems to be targeting the prostate," he says. "We suspect it's a combination of infectious agents and diet."

## CARCINOGENIC OOZE

De Marzo began to study inflammation in the prostate after noticing the strange pockets of shrivelled cells that he dubbed proliferative inflammatory atrophy (PIA). Despite their appearance, cells in PIA lesions proliferate at almost the same rate as cancer cells. Sometimes, PIA cells seem to merge with abnormal cells from regions of prostatic intraepithelial neoplasia (PIN), which are also thought to be a precursor to prostate cancer (see 'Cancer culprit'). And often, signs of chronic inflammation lurk nearby. "It looks like the inflammation might come first, and these lesions can result," De Marzo says.

Inflammatory cells can elicit the production of DNA-damaging oxidants. They also secrete the signalling proteins cytokines, which have an important role in regulating surrounding cells and can cause them to proliferate, De Marzo says. In other words, there are signs of oxidative stress, genetic instability and runaway cell division in areas where PIA, PIN and inflammatory cells huddle. "You're setting up the primordial ooze for carcinogenesis," says Lucia.

But not all inflammatory cells fight for team cancer. Some prevent precancerous lesions from taking hold. Researchers still have a long way to go to understand which cells, or combinations of cells, are helpful and which cause harm.

Lucia is focusing on a cytokine known as growth differentiation factor 15 (GDF-15), which is involved in regulating inflammatory pathways. GDF-15, he says, has been shown to slow the growth of tumours in the colon in animal studies. With James Lambert, a pathologist also at the University of Colorado's Anschutz Medical Campus, Lucia found that whereas GDF-15 was common in healthy prostate tissue, it was sparse in samples with chronic inflammation<sup>5</sup>. He suspects that the protein acts as a brake on inflammation — a useful tool for a gland that is situated so close to the urethra and the potential pathogens it harbours. "It could be that if

GDF-15 is inhibited, chronic inflammation develops," he says. Lucia is now exploring how GDF-15 might inhibit the tumour-promoting factors produced by some inflammatory cells, and possibly help to prevent prostate cancer.

Sfanos, meanwhile, is moving in a different direction by homing in on inflammatory cells that might increase the risk of developing prostate cancer — a daunting task. She is attempting to count and map the locations of different types of inflammatory cell in the prostate, starting with those that are known to be associated with other cancers. Eventually, Sfanos hopes that her work will reveal which combinations of cell types are harmful and which might be protective. "We hope to understand what is a good mix of inflammatory cell types versus what seems to be not so good, as far as the development of advanced disease," Sfanos says.

Physicians might then be able to run a test that determines the mixture of immune cells that are present in the cancerous prostate of a patient. "If there are more inflammatory cells of a certain type, or more immune cells in general, does that give us information about prognosis?" asks Platz. "If it does, perhaps those men need more or less surveillance going forward."

Such work on inflammation could have important implications for the prevention of prostate cancer. "We don't think it's a normal process for the prostate to grow too large or to get cancer," De Marzo says — a point of view that he acknowledges is counterintuitive, given the frequency of these conditions. "If it turns out there is an infectious cause — or two or three or four — and we could treat those, that might ultimately prevent a lot of disease." ■

**Kirsten Weir** is a freelance science writer in Minneapolis, Minnesota.

- Gurel, B. et al. *Cancer Epidemiol. Biomarkers Prev.* **23**, 847–856 (2014).
- Ravich, A. & Ravich, R. A. *New York State J. Med.* **51**, 1519–1520 (1951).
- Sutcliffe, S., Giovannucci, E., Isaacs, W. B., Willett, W. C. & Platz, E. A. *Int. J. Cancer* **121**, 2688–2692 (2007).
- Nakai, Y., Nelson, W. G. & De Marzo, A. M. *Cancer Res.* **67**, 1378–1384 (2007).
- Lambert, J. R. et al. *Prostate* **75**, 255–265 (2015).



Surgeon Declan Murphy positions a robotic device above a patient's abdomen so that a 3D telescope and surgical instruments can be installed.

## Q&A: Declan Murphy

# A robot convert

*In 2004, surgeon Declan Murphy was not convinced that using a robot to remove a cancer-riddled prostate was a significant improvement on keyhole, or laparoscopic, surgery. Eight-hundred robotic procedures later, he has not only changed his mind, but is now director of Robotic Surgery at the Peter MacCallum Cancer Centre in Melbourne, Australia.*

### How does robotic surgery compare with other surgical methods for removing a cancerous prostate?

There are some benefits to robotic radical prostatectomy over open surgery that are difficult to argue with. First, men can leave hospital much quicker: 85% of patients go home the next day. Second, the blood transfusion rates are significantly lower than for open surgery. Third, general surgical complications such as clots and infections also seem to be lower — and that is because it is minimally invasive surgery, like laparoscopic surgery.

Conventional laparoscopic prostatectomy, with a 2D view and straight instruments, is technically challenging, and this leads to longer operative times. A key paper published a few years ago (A. J. Vickers *et al. Lancet Oncol.* **10**, 475–480; 2009) showed that the learning curve for laparoscopic prostatectomy was very

long, much longer than for open surgery. The authors reported that around 750 cases were needed — which is a lifetime's work for many surgeons — before you would achieve your lowest cancer recurrence rates.

### How difficult is it to train with the robotic system?

The learning curve is much shorter than for laparoscopic prostatectomy — my colleagues and I estimated it was upwards of 80 cases. The device has some fantastic training features, such as a dual console, so it is like learning how to drive a car. There is also a touch screen that consultants can draw on so that our annotations will come up on the robot console, showing the trainee surgeon where to cut and where not to cut. However, robotic radical prostatectomy is a complex procedure that requires modular training and should only

be done by specialists. My department has an extremely strict list of requirements, and we frequently deny people access because they don't have the credentials.

### Why did you switch from laparoscopic to robotic surgery?

I was sceptical about the robot when I first had experience of it as a urology trainee at Guy's Hospital in London in 2004. I was of the opinion that you don't need a robot to do these operations, you just work harder and train harder with laparoscopic tools. But my view changed in 2007 when I undertook fellowship training in Melbourne, and I began to see data regarding outcomes of robotic surgery emerge. I realized that I would be able to achieve much better results for my patients by performing robotic prostatectomy rather than conventional laparoscopic or open surgery. You don't

ALAN MOYLE



ALAN MOYLE

suddenly become a fantastic surgeon just by using this device though. The surgeon's training and experience count for more than whether he or she is using the robot.

### What are the advantages of robotic surgery from the surgeon's perspective?

It is impossible to overstate how good the view is looking into this machine. The prostate is deep down in the pelvis behind the pubic bone, so it is difficult to get good views with open surgery, especially as there is more blood loss with this type of procedure. We have got used to very good views with laparoscopic surgery, but these are 2D — it's like having one eye closed when you are trying to stitch. With the robotic device, you are seeing in 3D, with a greatly magnified view in very high definition, which is unsurpassed by other approaches.

The other big advantage is the range of movement of the instruments. With laparoscopic surgery, we have straight instruments that do not have a 'wrist' on them. But suturing is a very dexterous movement. The robotic system has wristed instruments: you can turn your hand in the machine and the needle turns — a much more intuitive interface.

*"You don't suddenly become a fantastic surgeon just by using this device."*

### Does robotic surgery mean better outcomes for patients with cancer?

The problem with prostate cancer is that the outcomes take quite a number of years to materialize. The short-term surrogates for measuring cancer outcomes are things like positive surgical margins — when cancer cells are found right to the edge of the surgically removed tissue. If you have a positive surgical margin, you are five times more likely to need additional cancer treatment, such as radiotherapy, over the following two years. When we looked at data on 2,300 radical prostatectomies, we found a statistically significant 31% reduction in the number of patients with positive surgical margins after robotic prostatectomy.

We have also shown that there is a dramatic reduction in hospital stay after robotic surgery: from five days with open surgery down to just over one day. Furthermore, the blood transfusion rate for open surgery is 15%, whereas it's practically 0% with robotic surgery.

There are, however, two other important areas for patients undergoing radical prostatectomy where we cannot claim that robotic surgery is clearly better: urinary continence recovery and sexual function recovery. These are major quality of life outcomes that are very important to patients — and it is not possible to say with any confidence that robotic surgery

is any better than good open surgery by an experienced surgeon.

I get many patients who have had a biopsy taken or been offered open surgery and who are seeking a second opinion. I tell them that if they have come from a high-volume surgeon then, apart from the short-term outcomes of hospital stay length, blood transfusions and maybe margin rates, their longer term cancer outcomes are going to be just as good with those performing open surgery as they would be with us. But the reality is that in many regions today, Australia included, most fellowship-trained, high-volume surgeons, are using the robot and the amount of surgeons who have performed a large number of open procedures is dwindling.

### Have there been any randomized clinical trials comparing the types of surgery?

This is the major criticism over the years — we have failed to do randomized controlled trials. One such trial comparing robotic and

High-definition 3D images allow precise dissection.

open radical prostatectomy that has successfully recruited all of its patients is in Brisbane,

Australia, but a report is not expected until early 2016. The Brisbane trial, however, is the exception, and in many respects the boat has sailed. Hundreds of thousands of robotic procedures have already been reported in the literature in observational retrospective series, so everyone has already read about them and made up their mind. It is now almost impossible to sell a randomized controlled clinical trial to patients, or indeed to surgeons. We all know robotic surgery is better from a technical point of view and for the other short-term outcomes, so nobody wants to have open surgery any more.

### What are the downsides of robotic surgery?

The massive issue is the cost of the machine. It is made by a monopoly provider that has fiercely protected its patents — as it is entitled to do. The machines cost AUS\$2–3 million (US\$1.4–2.1 million), and there are also recurring costs; maintenance is about AUS\$250,000 per year and the surgical instruments we use cost AUS\$3,500 per operation. They are reusable, but only up to 10 times.

There is a practical difficulty as well. Although you have fantastic vision and magnification, there is no tactile feedback from the wristed instruments — you can't feel anything — and surgery has in the past relied heavily on sense of touch. However, the greatly superior vision more than makes up for this.

### Are the costs of robotic surgery balanced by the benefits?

The costs of the machine can be offset by reductions in the length of hospital stay and number of blood transfusions, and there is a critical number where it becomes cost effective. In our model, that number is 140 cases. If you're amortizing a AUS\$3 million device over 7 years, including an annual maintenance contract, a really important part of diluting the cost is to have a high volume of surgery on the machine.

Radical prostatectomy numbers are decreasing; the reason is not to do with the robot cost, however, but with changes in early prostate-cancer screening, detection and patterns of care. The number of men being offered or asking for a prostate-specific antigen test has dramatically dropped, so the first reason why fewer radical prostatectomies are being done is because fewer men are being tested. Another reason for the decline is the rise in active surveillance as a management option for early prostate cancer (see page S126).

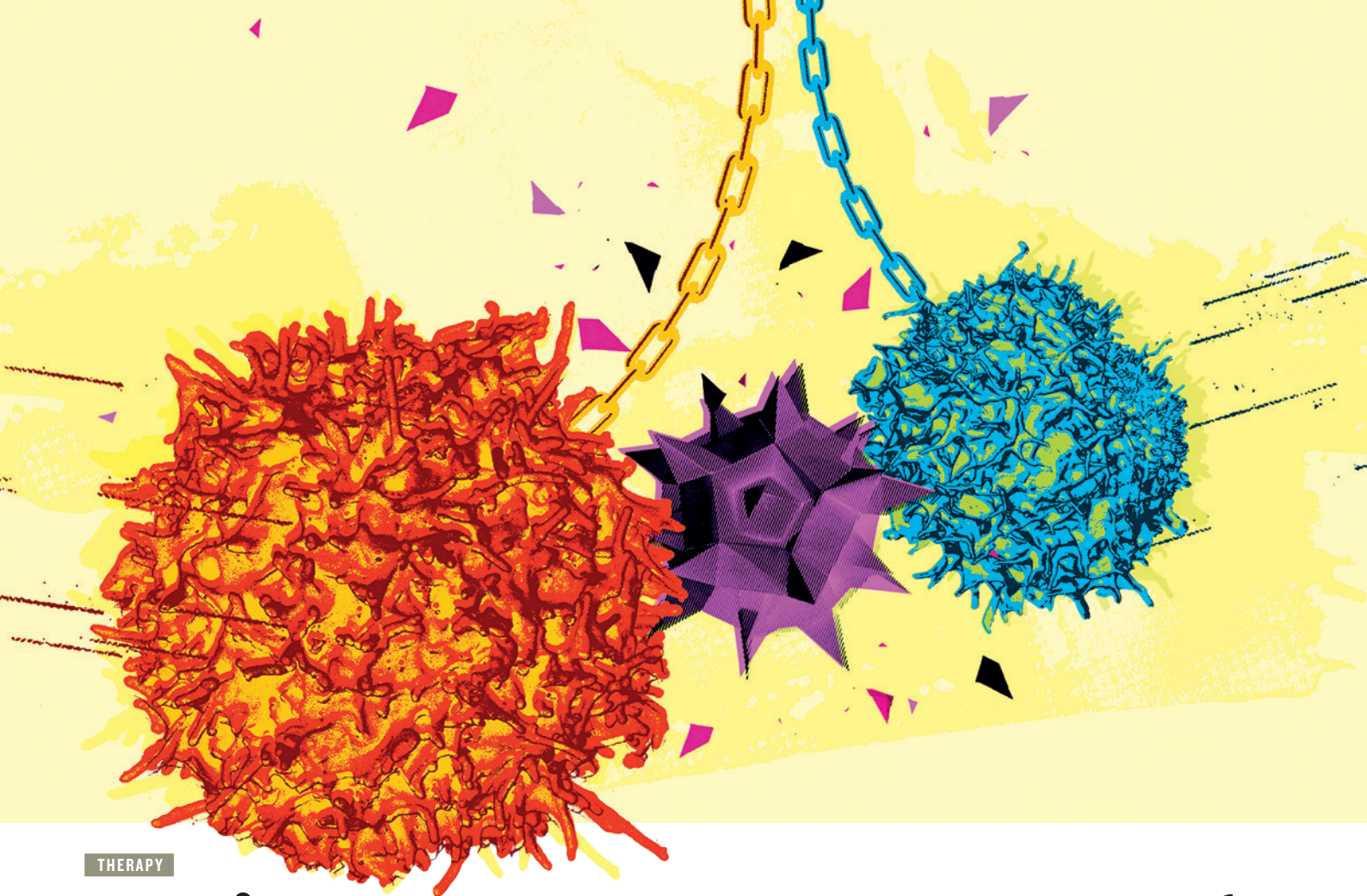
INTERVIEW BY BIANCA NOGRADY

This interview has been edited for length and clarity.



ALAN MOYLE





THERAPY

# An immune one-two punch

*Combination therapies that activate the immune system in complementary ways could help more men with prostate cancer to contain their disease long term.*

BY KATHERINE BOURZAC

When cancer immunologist Michael Curran was a postdoc, he made a discovery of the magnitude that scientists only dream about. He showed that two antibodies that unleash the immune system had a synergistic effect, bringing about the eradication of melanoma tumours in mice. What is more, this effect also worked in people. Curran and his colleagues published their mouse results in 2010 (ref. 1); subsequent clinical trials showed that the combination therapy is so effective at treating people with melanoma that some patients are “durably cured” of their cancer, he says.

Immunotherapy works well for people with melanoma, and researchers such as Curran, now an immunologist at the University of Texas MD Anderson Cancer Center in Houston, are trying to create similarly dramatic effects in other cancers. But Curran’s therapy does not work for prostate cancer — not even in mice. Immunotherapies are new, and researchers are still figuring out how

they work, says Curran. There is one immunotherapy for prostate cancer approved for use, and only in the United States. Sipuleucel-T adds, on average, a few months to a man’s life. But anecdotally, oncologists report men who have undergone the therapy living for years without needing further treatment.

To make prostate-cancer-immunotherapy success stories more common, physicians and immunologists need to understand why some men respond to the treatment, and some do not. Such insights will help them to predict which patients are most likely to benefit from these expensive treatments, and could guide the design of new versions that work better for more people. Several clinical trials are testing cancer vaccines (see ‘Immunotherapy on trial’), as well as therapies that combat a tumour’s tendency to muffle immune responses. Many of these trials are exploring combinations of therapies that act on different immune-system or cancer pathways, to

make sure that tumour-killing T cells are fully equipped to do their work.

## CHASING THE LONG TAIL

For a patient whose prostate cancer has spread to the lungs, bone or elsewhere, the prognosis is bleak. Chemotherapy and radiation shrink tumours and extend life by a few months, but then they stop working — either because the tumour mutates to get around a targeted therapy or because patients are taken off the treatments because of the side effects. Immunotherapy drugs can have longer term effects when they work well, but so far that is rare for prostate cancer.

Sipuleucel-T is controversial. The median survival benefit is only four months<sup>2</sup> — about the same as conventional therapies — and it costs US\$93,000. That kind of limited benefit and high cost is not unheard of for cancer drugs in the United States, but it is unusual. And sipuleucel-T is more complicated to administer than a conventional drug. Unlike most drugs, which come premade and can be sold off-the-shelf, sipuleucel-T is

GARY NEILL

► **NATURE.COM**

Read more on cancer immunotherapy at:  
[go.nature.com/kvpvgzl](http://go.nature.com/kvpvgzl)



personalized. The patient's white blood cells are separated from their blood and sent to a central processing facility. There, these cells are incubated with the enzyme prostatic acid phosphatase — to train them to seek out cancer cells that produce this protein. The cells are then returned to a local clinic and infused back into the patient. This process is done three times. And although the cell harvesting can be performed at any Red Cross blood bank in the United States, it is still much more complicated than writing a prescription for a pill or sending a patient to an infusion clinic for conventional chemotherapy, says Lawrence Fong, an immunologist who treats men with prostate cancer in his clinic at the University of California, San Francisco.

That complexity, and the expense that accompanies it, have brought about a backlash against sipuleucel-T. The therapy's creator Dendreon, based in Seattle, Washington, received US Food and Drug Administration approval to market sipuleucel-T in 2010. But when the drug's poor sales figures were revealed just a year later, the company's stock plunged 67% in a single day. In November 2014, Dendreon filed for bankruptcy; its assets were sold off the following February.

Montreal-based Valeant Pharmaceuticals, who picked up the drug, withdrew an application to market it in the European Union in May 2015. Questions were raised about the expense of the treatment and about the clinical trials, says Hardev Pandha, an oncologist at the University of Surrey, UK. "There were a few infusions and then that was it," he says. "It wasn't seen as a sustained treatment."

One factor weighing against broader acceptance is that the mechanism underpinning how sipuleucel-T works in patients was not established in the initial clinical trials. It is thought to work with dendritic cells in the blood. These cells have receptors that recognize the chemical signatures of microbes, cancer cells and other antigens, and when they spot one, the dendritic cells attach it to a protein on their surface like a red flag. These warnings kick-start T cells into action, spurring them to hunt down and kill foreign cells that display the antigen. But the clinical trials did not look for activated T cells or their markers in patient samples, says urologist Martin Sanda at Emory University School of Medicine in Atlanta, Georgia, and for that reason it is difficult to know why it works for some men and not for others. Researchers are now investigating the activity of specific cells in men who have had the treatment.

That the complex therapy works very well in some men is reason enough for many physicians to offer it. "I have to advocate for my patients," says Fong. He and other oncologists know that some patients respond well to immunotherapy — something that is not reflected in the average survival numbers. Their tumours do not shrink, but they stop



Administering sipuleucel-T to patients is much more complicated than providing conventional therapies.

growing, and some men are stable for years.

Survival rates for patients with late-stage disease who are given conventional treatment plunge to zero after a year or two. By contrast, the graph of survival over time for those given immunotherapy has a 'long tail' — never reaching zero in clinical trials. For Fong, one patient in particular illustrates the hope for the therapy. The patient's recurrent metastatic prostate cancer had become resistant to hormone therapy. "He got the usual treatment and responded as most patients do," Fong says. The treatments work for a while, shrinking tumours for a few months, after which they grew anew. Then Fong treated him with a course of sipuleucel-T. Five years later, his cancer has not grown, nor has he needed further treatment.

## CANCER VACCINES

Even those such as Fong who offer the treatment to their patients agree with Pandha, who says that immunotherapy needs to move away from "bespoke personalized medicine" like sipuleucel-T. To that end, researchers are working on off-the-shelf vaccines for prostate cancer. Furthest along in clinical trials is a vaccine developed at the US National Cancer Institute (NCI).

Called PROSTVAC, this therapy borrows from the playbook of infectious diseases, using two weakened viruses — vaccinia and fowlpox — engineered to carry prostate-specific antigen (PSA). The vaccine has been in the works since the late 1990s, starting in the lab of NCI immunologist Jeffrey Schlom. It showed promising results in phase II clinical trials, in which patients remained progression free for an average of 12 months<sup>3</sup>, and it is now in phase III clinical trials for treating metastatic prostate cancer.

Vaccines target specific antigens — and in the case of PROSTVAC, PSA is the molecule of choice. PSA is a self-antigen: it is made by healthy, as well as cancerous, prostate tissue. But PROSTVAC is a therapeutic vaccine,

which is intended to be given only to men who have already had their cancerous prostate gland removed. In these men, the only cells producing PSA — and therefore the only cells that the vaccine will target — are cancer cells. The vaccines also seem to have an effect called antigen spreading, says James Gulley, a tumour immunologist at NCI. Once the immune system identifies and attacks the tumour, it recognizes and goes after other tumour antigens that it finds on its own.

Researchers have discovered additional targets for treating prostate cancer, and some believe that the immune system may be able to mount a better response to vaccines that target an antigen that is unique to the tumour, rather than a self-antigen such as PSA — or to one that targets multiple antigens.

Many transcription factors — regulatory proteins that promote or block gene expression — are overexpressed in tumour cells and so are a good target for cancer therapy. Sanda is testing, in animal models, whether the transcription factors ERG and SIM2 can be used as antigens.

Charles Drake, an oncologist and immunologist at the Johns Hopkins School of Medicine in Baltimore, Maryland, is taking a different approach: a quadruple-antigen vaccine akin to a Swiss Army knife. This experimental vaccine uses prostate acid phosphatase — the same antigen used in sipuleucel-T cell therapy — along with another protein called prostate-specific membrane antigen. Both are found in normal prostate tissue, but a third antigen is specific to prostate cancer. And a fourth is a protein that is overexpressed in cells left behind after prostate removal, and is considered a prostate-cancer precursor gene product.

Instead of a virus, Drake's vaccine uses attenuated *Listeria* bacteria as the carrier. These weakened microbes have been used in other vaccines, including one for pancreatic cancer, which Drake says elicited a strong immune response in a phase II trial led by another group at Johns Hopkins. He hopes to see the

## THERAPY

*Immunotherapy on trial*

After disappointing results from the first approved immunotherapy for prostate cancer, researchers are developing a host of alternatives that could deliver better results for more patients.

**PROSTVAC (phase III).** Developed at the National Cancer Institute, this multicourse viral vaccine activates the immune system against prostate-specific antigen.

**Hormone and checkpoint therapy (phase II).** The high levels of testosterone in prostate tumours inhibit the activity of cancer-killing T cells. Combining hormone therapy with the checkpoint therapy ipilimumab could combat this.

**PROSTVAC and ipilimumab (phase II).** By combining a viral vaccine and a checkpoint therapy, it is hoped that one will activate

T cells and the other will keep the tumour from suppressing them.

**Sipuleucel-T with checkpoint therapy and chemotherapy (phase II).** This trial is looking for synergy between the cell therapy sipuleucel-T and checkpoint therapy, along with the conventional chemotherapy drug cyclophosphamide.

**Sipuleucel-T with ipilimumab (phase II).** By combining sipuleucel-T with checkpoint therapy researchers hope to determine what order they should be given in — block tumour suppression first then provide cell therapy, or the other way around?

**DNA vaccine (phase II).** Instead of a microbial carrier, this vaccine uses a naked DNA plasmid that codes for prostate acid phosphatase. **K.B.**

same in the first clinical trials of the prostate-cancer vaccine, set to begin by early 2016.

Drake says *Listeria* is easier to grow in culture than vaccinia and fowlpox. And the *Listeria* vaccine can be given multiple times without the need for different carriers like PROSTVAC. Other researchers are experimenting with vaccines that use DNA, with no carrier at all. A phase II trial of a DNA vaccine now under way will indicate whether this method elicits as strong an immune response as the attenuated pathogens — a result that will be of keen interest to researchers such as Sanda who have not yet chosen a carrier for their novel antigen targets.

Trials of these vaccines depend on better monitoring of biomarkers, which are the key to finding out why some patients respond very well and others not at all. “We’re learning how to collect patient samples and not just look at everything in a mouse,” says MD Anderson oncologist Padmanee Sharma. Although much can be learned from mouse studies, she says, the interconnected co-evolution of tumour and immune system needs to be studied in people.

**COMBO DEAL**

Tumours take advantage of naturally occurring checkpoints that prevent healthy immune reactions from becoming dangerous. Once a T cell is activated by an antigen and expands its numbers, it starts expressing a checkpoint receptor. “This puts an expiration on T cells of a day or three,” says Curran. That is a good thing — you do not want billions of killer immune cells accumulating in your lungs after your cold clears up. But tumours turn this safety mechanism to their advantage, using that receptor as a target for its own suppressive signals, so that T cells never get going.

Checkpoint therapies target and block these suppressive signals.

T cells have to be attracted to a tumour in the first place, however, otherwise checkpoint therapy has no effect. Treatments such as vaccines and cell therapies (sipuleucel-T) stimulate this process, and a combination of these treatments and checkpoint therapy maybe the best way forward. This is now being tested in patients.

The conventional wisdom is that checkpoint therapy works well in mutation-rich cancers

*“We’re learning how to collect patient samples and not just look at everything in a mouse.”*

such as melanoma because these tumours generate high levels of novel antigens that attract T cells to the tumour. The T cells then just need a little boost from checkpoint therapy. Prostate tumours are not as rich in T cells, a deficit that researchers suspect is because they have too few mutations to catch the immune system’s attention.

Curran thinks it is more complicated than that. After all, he says, prostate cancer is not uniquely low in mutations among cancers — it ranks somewhere in the middle. On average, prostate cancers have about 50 mutations — and each one of them should be read by the immune system as an antigen. That is almost five times as many potential antigen targets as the influenza A virus, which does provoke an immune response. It is not just about the numbers, Curran argues.

When Curran saw how much better checkpoint therapy worked against melanoma than prostate cancer, he came up with a new approach. He concentrated on the tumour microenvironment. Prostate tumours differ

from healthy tissue in that they contain high levels of testosterone and low levels of oxygen. “That’s everything T cells hate,” he says. Besides which, the tumours are poorly vascularized — they are a backwater on the circulatory system that the T cells travel.

In 2011, Curran recalled something he had heard in graduate school about drugs that target tumour hypoxia, and wondered if that may be an avenue to improve the effectiveness of immunotherapy. To explore that possibility, he began a collaboration with Threshold Pharmaceuticals in South San Francisco, California, which makes a drug called evofosfamide. This compound circulates in a non-toxic form until it reaches a region of low oxygenation, which triggers the release of a DNA-damaging agent. Curran wondered what would happen after the drug had killed tumour cells in the hypoxic areas of tumours. Would those areas become a wasteland — or would T cells find a foothold? Curran found that, in a mouse model of prostate cancer, cancer-cell killing is followed by a wound-healing response and the growth of new blood vessels. That brings oxygenated blood and, it seems, a more T-cell friendly environment. “T cells can then enter the areas they were formerly blocked out of,” says Curran. After introducing the evofosfamide, Curran and his colleagues in Houston administered checkpoint therapy to prevent the arriving T cells from being suppressed. Curran reported these results at The Inaugural International Cancer Immunotherapy Conference this year and is now designing a human trial of this combination.

Hypoxia is not the only environmental barrier to T cells. Another combination-therapy clinical trial addresses the high levels of immune-suppressing testosterone in prostate tumours by combining hormone therapy (to lower testosterone) and checkpoint therapy. And it is now becoming evident that some chemotherapies that were thought to work only by killing cancer cells are dependent on the immune system to work. They might also be fruitfully combined with checkpoint therapy to fight prostate cancer.

Combination therapy is the great hope of prostate-cancer immunologists. There is no guarantee that these therapies will avoid the cost problems associated with sipuleucel-T. But if combining treatments allows more men to go into remission — or perhaps even be cured — the high price tags may not raise as many eyebrows. For immunotherapy, says Pandha, combinations are “the final piece of the jigsaw.” ■

**Katherine Bourzac** is a freelance science writer in San Francisco, California.

1. Curran, M. A., Montalvo, W., Yagita, H. & Allison, J. P. *Proc. Natl. Acad. Sci.* **107**, 4275–4280 (2010).
2. Kantoff, P. W. et al. *N. Engl. J. Med.* **363**, 411–422 (2010).
3. DiPaola, R. S. *Eur. Urol.* **68**, 365–371 (2015).

# nature INDEX 2015

CHINA





*Despite advances in detection and therapy, much about this common malignancy remains unknown. Here are some of the most important unresolved issues.*

BY RICHARD HODSON

## PROSTATE CANCER

# 4 BIG QUESTIONS

### QUESTION

### WHY IT MATTERS

### WHAT WE KNOW

### NEXT STEPS

1

**What causes prostate cancer?**

Worldwide, prostate cancer is the second most common cancer in men, after lung cancer. Identifying a preventable cause of this disease could reduce the number of cases.

Risk increases with age, and inherited factors are estimated to be responsible for 5–9% of cancers. Risk is five times higher in men with *BRCA2* mutations. Despite extensive research, the disease has not been clearly linked to any preventable risk factors.

Taking into account the differing rates of prostate-specific antigen (PSA) testing in populations could help to firm up links. Arsenic and cadmium compounds, anabolic steroids and ionizing radiation may be causes; carrots and soya may reduce the risk.

2

**Is PSA testing an effective method of screening?**

Measuring levels of PSA in the blood is often used to detect prostate cancer. Without a reliable test, in some cases, the first symptoms of the cancer are signs that it has spread to the bones, where it is much less treatable.

Rates of diagnosis spiked in the 1990s in the United States, partly because of the use of PSA screening for men without symptoms. But it is likely that many men were unnecessarily treated for cancers that would never have caused harm.

The PSA test could be a useful procedure if it is applied using evidence-based guidelines (page S123). Combining the screen with other analyses, such as testing for genetic markers, could reduce the number of unnecessary treatments.

3

**Is it safe to leave prostate cancer alone?**

The most common treatments for localized disease — removal of the prostate and radiotherapy — have side effects, such as incontinence and sexual dysfunction. Men with less-aggressive tumours might be better off avoiding these procedures.

Half of US men with low-risk prostate cancer between 2010 and 2013 had their prostates removed, whereas 40% opted to watch and wait. Some studies have suggested that low-risk patients can be safely monitored for more than a decade.

The challenge facing active surveillance is knowing which men have slow-growing tumours that can be left, and which are more aggressive. New methods for telling aggressive and indolent cancer cells apart are being investigated.

4

**Can advanced prostate cancer be treated?**

Once prostate cancer has spread to the lymph nodes and bones, the outlook is poor. Five-year survival rates for metastatic cancer are one-third of those for localized disease; advanced prostate cancer is considered incurable.

Therapies for advanced prostate cancer have emerged only in the past decade. The go-to treatment is chemical castration: drugs are used to suppress male hormones. This can prolong life by two or three years before tumours become resistant.

Drugs designed to treat castration-resistant tumours are also facing a resistance problem — 20–40% of patients do not respond to these therapies, and their efficacy is eventually lost in all men (see page S128).

Richard Hodson is supplements editor for Nature.



# nature INDEX 2015

## CHINA

NATURE, VOL. 528, NO. 7582 (DECEMBER 17, 2015)

COVER ART: ALISDAIR MACDONALD

A year has passed since we published the first Nature Index supplement about China, and the data accrued in that time reveal another remarkable period for science in that country. In the past 12 months, the growth of China's output in the index has dwarfed that of any other nation.

In this supplement, we analyse three years of research output from China — from 2012 to 2014 — providing a telling snapshot of the country's emergence as a scientific superpower, a phenomenon watched with intense interest around the globe. The articles in this index focus on cities with particularly interesting stories to tell.

Nationally, China's weighted fractional count (WFC) rose 37% between 2012 and 2014, and growth in this metric was notably high in Hangzhou, Xi'an and Chengdu (see 'At the very heart of progress', S179). Further explanation of how WFC and other metrics are calculated can be found in the guide to the index on page S190.

Measuring output by WFC reinforces the status of Beijing, Shanghai and Nanjing as the dominant scientific centres, which our feature article on page S176 explores.

In light of China's ongoing drive to use science and technology to move away from its economic reliance on traditional manufacturing, we also examined the nation's industry hubs.

The index shows that cutting-edge life science has matured quickly in Shenzhen, Beijing and Wuhan. In these cities front-line science is yielding practical outcomes and bringing returns that will stoke the fires of the Chinese economy. The city of Shenzhen, in particular, has experienced a remarkable transformation into a research-based industry hub and companies based there now account for almost half of the country's international patent filings (see 'The changing face of industry', S184).

As China cements its role as the world's second largest producer of high-quality research papers, it is gaining momentum from academic collaborations. Nature Index data reveal that Hong Kong, Hefei and Tianjin are active in the pursuit of international or domestic associations. All three recorded a high collaboration score, a metric of institutional collaboration in terms of co-authorship of articles in journals covered by the index.

Each year, the Nature Index presents a more comprehensive picture of the patterns driving research. China's scientific ascension is likely to continue, a phenomenon that the index, and future China-focused supplements, will be well placed to follow.

**Karen McGhee**  
*Guest editor, Nature Index*  
**Nicky Phillips**  
*Senior editor, Nature Index*

## CONTENTS

### S166 BUILDING A POWERHOUSE

Cities across the land are making a mark in every capacity

### S170 THE RAPID RISE OF A RESEARCH NATION

Huge increase in quality research parallels the economic boom

### S176 THREE GIANTS TIGHTEN THEIR GRIP

Beijing, Shanghai and Nanjing continue their dominance

### S179 AT THE VERY HEART OF PROGRESS

Chengdu, Hangzhou and Xi'an each find their own niche

### S184 THE CHANGING FACE OF INDUSTRY

Research leads to bold new horizons for manufacturing hubs

### S187 ALLIANCES FOR SCIENTIFIC SUCCESS

Partnerships old and new reap rich rewards of collaboration

### S190 HOW TO USE THE INDEX

A guide to getting the most out of Nature Index data

### S191 THE TABLES

How China's many institutions stack up

**EDITORIAL:** Stephen Pincock, Nicky Phillips, Karen McGhee, Rebecca Dargie, Victoria Kitchener, David Cyranoski, Hepeng Jia, Sarah O'Meara, Peng Tian, Yingying Zhou. **ANALYSIS:** Larissa Kogeleck. **ART & DESIGN:** Alisdair Macdonald, Kate Duncan, Chris Gilloch. **WEB & DATA:** Bob Edenbach, Olivier Lechevalier, Naomi Nakahara, Pamela Sia, Jörn Ishikawa, Yuxin Wang, Jyoti Miglani, Jennie Pao, Akiko Murakami, Takeshi Ouchi. **PRODUCTION:** Sue Gray, Karl Smart, Ian Pope, Chandler Gibbons, Mira Loutfi, James McSweeney. **MARKETING:** Hannah Phipps. **SALES:** Janet Cen, Stella Yan, George Sun. **ART DIRECTOR:** Kelly Buckheit Krause. **PUBLISHING:** Nick Campbell, Richard Hughes, David Swinbanks.

#### NATURE INDEX 2015 CHINA

The Nature Index 2015 China, a supplement to *Nature*, is produced by Nature Publishing Group, a division of Macmillan Publishers Ltd. This publication is based on data from the Nature Index, a website maintained by Nature Publishing Group and made freely available at [natureindex.com](http://natureindex.com).

Nature Editorial Offices  
The Macmillan Building  
4 Crinan Street,  
London N1 9XW, UK  
Tel: +44 (0)20 7833 4000  
Fax: +44 (0)20 7843 4596/7

#### CUSTOMER SERVICES

To advertise with the Nature Index, please visit [natureindex.com/support](http://natureindex.com/support)  
[feedback@nature.com](mailto:feedback@nature.com)  
Copyright © 2015 Nature Publishing Group.  
All rights reserved.

# BUILDING A POWERHOUSE

The story of China's phenomenal growth in scientific output during the past three years can be told through the experience of eleven cities. Each has displayed impressive output, measurable in one way or another, as determined by analysis of Nature Index data from 2012 to 2014.

Four index metrics have been used to evaluate the performances of China's cities: article count (AC); fractional count (FC); collaboration score and weighted fractional count (WFC). (For a full explanation of these metrics see S190.)

Represented here in yellow are China's scientific heavyweights Beijing, Shanghai and Nanjing: the cities that have shown the highest total output. The index data also reveal some of the cities where total scientific output has been growing fastest — China's rising stars — are Xi'an, Chengdu and Hangzhou, shown in orange. Delving deeper into index data identifies Shenzhen, Beijing and Wuhan,

represented in red, as the nation's industrial research powerhouses; where high scientific output is being used to generate economic return. Meanwhile, the cities most actively pursuing partnerships to advance scientific discoveries are Hefei, Tianjin and Hong Kong, in purple. ■

Data analysis by Larissa Kogleck

## ARTICLE COUNT

These 11 cities generated

72%

of China's overall articles in the Nature Index in 2014.

BEIJING AC: 5,163

SHANGHAI AC: 1,955

NANJING AC: 1,064

HEFEI AC: 696

WUHAN AC: 619

HONG KONG AC: 600

TIANJIN AC: 461

HANGZHOU AC: 458

XI'AN AC: 309

CHENGDU AC: 287

SHENZHEN AC: 165

1.4B

TOTAL ESTIMATED POPULATION OF CHINA, THE WORLD'S MOST POPULOUS COUNTRY.

## RISING STARS

Xi'an, Chengdu and Hangzhou experienced significant growth in their contribution to the Nature Index.

## XI'AN

>2014 WFC: 141  
>2013 WFC: 95  
>2012 WFC: 58

## CHENGDU

>2014 WFC: 131  
>2013 WFC: 107  
>2012 WFC: 73

2.1%

PERCENTAGE OF GDP SPENT ON R&D IN 2014.

8,641

CHINA'S ARTICLE COUNT 2014

6,037

WEIGHTED FRACTIONAL COUNT 2014

6,328

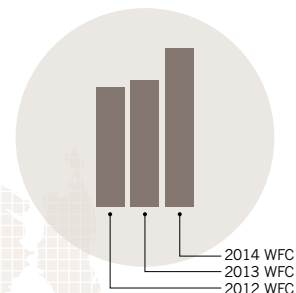
FRACTIONAL COUNT 2014

4,541

CHINA'S COLLABORATION SCORE 2014

## LEGEND

Circle size is relative to WFC in 2014.



## EXPLORE CHINA

Read more about the cities behind China's rapid scientific ascension.

- **STRONGHOLDS** ▶ **PAGE S176**
- **RISING STARS** ▶ **PAGE S179**
- **INDUSTRY HUBS** ▶ **PAGE S184**
- **COLLABORATORS** ▶ **PAGE S187**

**BEIJING**

>2014 WFC: 1,504  
>2013 WFC: 1,375  
>2012 WFC: 1,181

**STRONGHOLDS**

Beijing, Shanghai and Nanjing continue to dominate China's scientific output.

**TIANJIN**

>2014 WFC: 164  
>2013 WFC: 168  
>2012 WFC: 136

**HEFEI**

>2014 WFC: 255  
>2013 WFC: 212  
>2012 WFC: 179

**NANJING**

>2014 WFC: 388  
>2013 WFC: 304  
>2012 WFC: 286

**SHANGHAI**

>2014 WFC: 833  
>2013 WFC: 693  
>2012 WFC: 605

**WUHAN**

>2014 WFC: 257  
>2013 WFC: 217  
>2012 WFC: 192

**HANGZHOU**

>2014 WFC: 221  
>2013 WFC: 178  
>2012 WFC: 143

**SHENZHEN**

>2014 WFC: 54  
>2013 WFC: 36  
>2012 WFC: 28

**INDUSTRY HUBS**

Wuhan, Shenzhen and Beijing are home to corporations that are making a significant contribution to research in the Nature Index.

**HONG KONG**

>2014 WFC: 248  
>2013 WFC: 242  
>2012 WFC: 220

**COLLABORATORS**

Hong Kong, Hefei and Tianjin have made the most of collaborating with other institutions in China and the world.





China's many hopeful and determined graduates take their place in a rich and varied research landscape that is transforming the country's fortunes.

# THE RAPID RISE OF A RESEARCH NATION

*China's economic boom is mirrored by its similarly meteoric rise in high-quality science.*

BY YINGYING ZHOU

China has ambitious plans to source as much as 15% of its energy from renewable sources by 2020, at the same time its economy is projected to slow. It also aspires to be the next space superpower while facing major health and environment challenges, such as an ageing population and water shortages.

The Chinese government knows that surmounting these challenges while achieving its goals can only be accomplished through science. Indeed, China is pegging its future prosperity on a knowledge-based economy, underpinned by research and innovation. For a country that invented paper, gunpowder and the compass, such lofty ambitions could be realized. This year pharmacologist Tu Youyou became the first Chinese researcher to be awarded a Nobel Prize in Medicine for helping discover a new drug for malaria that has saved millions of lives.

"With a solid base built upon the large quantity of research, China is [about] to take off in world-leading innovations and scientific breakthroughs," says He Fuchu, the founding

president of PHOENIX, the Chinese National Center for Protein Sciences. "High-quality research is built upon the accumulation of incremental advances," He says.

The Nature Index shows China is already a high-quality scientific powerhouse. Since the first Nature Index database started in 2012, China's total contribution has risen to become the second largest in the world, surpassed only by the United States.

But, what sets China apart is the rapid growth of its WFC. While China's contribution grew 37% from 2012 to 2014, the United States saw a 4% drop over the same period.

## AN ECONOMIC IMPERATIVE

A key driver of China's scientific progress is its burgeoning economy, dazzling the world since embarking in the early 1980s on a transformation from a centrally planned to market-based economy. In the past three decades, China has achieved a consistently impressive annual average GDP growth rate of around 10%, and has overtaken Japan to become the world's second largest national economy behind the United States.

"The economic success has fuelled the

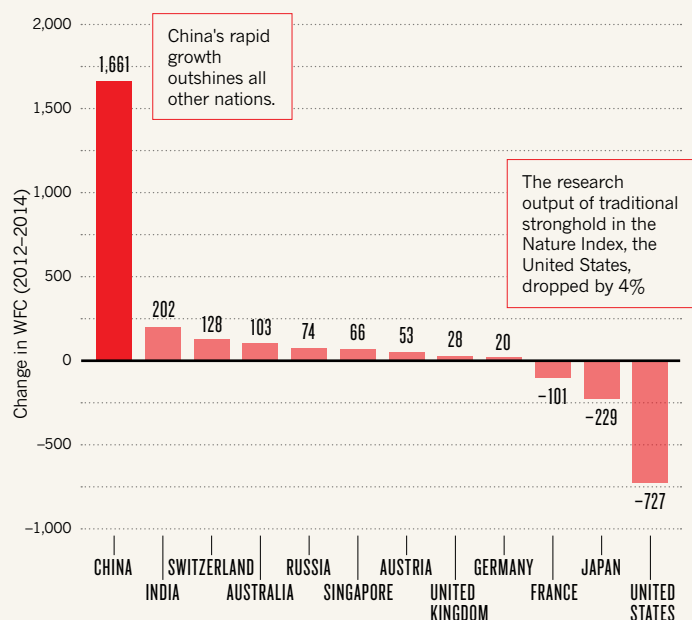
nation's investment in science and technology," says Liu Zhu, a researcher at Harvard University's Kennedy School of Government, who is also currently a fellow in sustainability science at the California Institute of Technology. While China's unfettered growth cannot last forever — economic growth has slowed, with the GDP growth rate falling to 7.4% last year, its lowest in 25 years — it has been the subject of global awe and fascination.

Figures from the National Bureau of Statistics of China show that during the past decade the nation's total research and development (R&D) expenditure also blossomed, achieving an average growth rate of more than 20%. In 2014, R&D expenditure totalled 1,330 billion yuan, equivalent to 2.1% of the national GDP. China's rapid growth trajectory in R&D investments is in sharp contrast to the constraints placed on R&D budgets of the United States, Japan, and most European countries, still recovering from the global economic crisis that began in 2007.

Recognizing the importance of scientific research in driving technological innovation and economic progress in 2006, the Chinese government unveiled its National

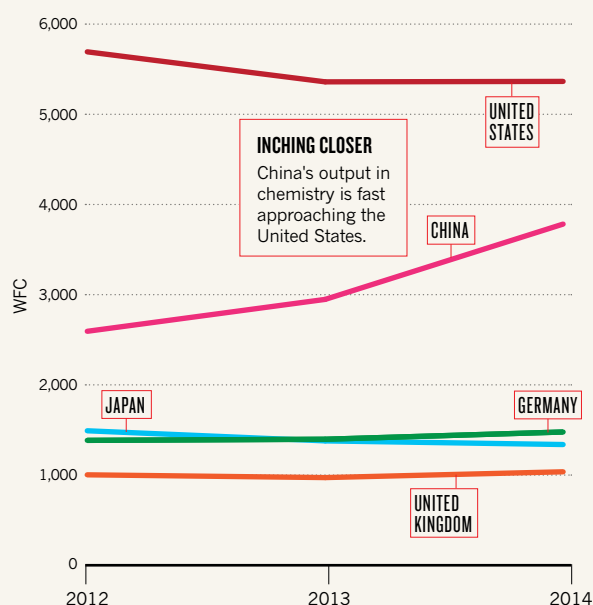
## CHINA BOOMING

This cross-section of countries in the Nature Index shows how remarkable China's increase in high-quality science output has been in recent years.



## CHEMISTRY CHAMPS

Change in output of the top five leading countries in chemistry.



Medium- and Long-Term Plan for the Development of Science and Technology, setting out a path to transform the country into a “science powerhouse” by 2020. The 15-year plan called upon an “indigenous innovation” campaign, putting science and technology development at the centre of the national development strategy. Under the strategy, investment in higher education was emphasized, recognizing that human resources are at the heart of scientific development.

China has made great efforts to expand its higher education system and enlarge its scientific workforce. The number of PhD graduates in science and engineering has soared in the past decade along with the number of graduates with bachelor degrees. Central and local government efforts to attract Chinese-born

*“China has the chance to be a research giant and establish a long-term culture of innovation.”*

scientists to return from overseas work and study have also paid off. The prominent 1,000 Talent Plan initiated in 2008 by the Central Organization Department of the Chinese Communist

Party has hugely exceeded its eponymous goal, having attracted more than 4,180 top-level scientists from abroad by mid-2014. The cumulative number of returned PhD holders reached 110,000 in 2014.

“The improvement of the research capabilities of Chinese researchers and the returning

of foreign-trained Chinese scientists from overseas certainly adds to the momentum of China’s scientific growth,” says Wang Jun, the former chief executive officer, and now partner, of successful genomics sequencing company BGI. David Reiner, a senior lecturer in science and technology policy from Cambridge University’s business school and a keen observer of China, says the return of its large scientific diaspora also fosters a supportive research culture.

With a deep cultural reverence for education, the Chinese hold scientists and their research in high regard. “The government investment and promotion, combined with



Wang Jun says returning scientists help growth.

the determination of Chinese scientists and societal support have fostered a culture of innovation,” says Wang. “China now has the opportunity not just to be seen as a global research giant but also to establish a long-term culture of innovation that will undoubtedly lead to myriad scientific discoveries in decades to come.”

## BOLSTERING QUANTITY WITH QUALITY

The research assessment system has also played a role in China’s rapid growth in output. An increasing focus on evaluation systems that measure quality is shifting emphasis from quantity-driven metrics. Most universities and research institutions now evaluate researchers based on the number of publications in high-impact journals rather than the sheer volume of publications, according to Ren Xiaobing, chairman of Xi’an Jiaotong University’s Frontier Institute of Science and Technology.

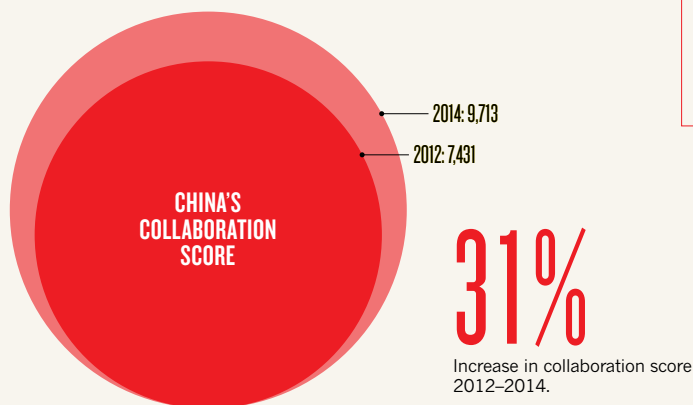
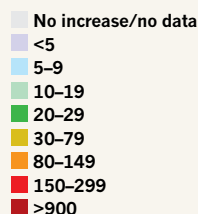
But Reiner warns that the increased pressure on researchers to focus solely on publishing their work in academic journals could encourage academic fraud. “The downside of an exclusive focus on quantitative metrics is that they may blind what is really important to research, making it difficult for scientists to explore blue-sky research ideas,” he says.

Some major research institutions, such as the Chinese Academy of Sciences (CAS) and prestigious universities, such as Xi’an Jiaotong University, are beginning to include other evaluations in their researcher assessments. “Other than the criterion of high-impact publications,

## COLLABORATION HOTSPOTS

This map shows countries that have experienced an increase in their collaboration score with China between 2012 and 2014. Collaboration score = sum of the fractional count (FC) for each of China's bilateral partnerships.

## ABSOLUTE COLLABORATION SCORE INCREASE



## UNITED STATES

The research output of China's partnership with the United States has dramatically increased its collaboration score from 3,714 in 2012 to 4,664 in 2014, a rise of 949.

## CANADA

The China–Canada collaboration experienced the third largest increase.

## CHILE

China's partnership with Chile has increased the most in this region.

## CHINA'S SUPERSTAR



A congress of the Chinese Academy of Sciences, a central plank of the country's research prowess.

The Chinese Academy of Sciences (CAS) is the world's largest institutional contributor to the Nature Index. In 2014 its WFC was 1,308 (its AC was 3,124), significantly higher than that of the second-ranked institution, Harvard University, with a WFC of 865. By subject areas, CAS leads not only in chemistry, but also in physical sciences, and earth and

environmental sciences, with higher WFCs in these major subject areas than any other research institutions worldwide. CAS employs more than 60,000 people, has 104 research institutes and has been central to China's modern scientific development. Its 2015 budget was 54 billion yuan (US\$8.4 billion), a 9% increase from 2014. ■

we have also adopted more nuanced and individual-focused assessment criteria that are based on a researcher's relative performance compared to his or her peers," says Ren.

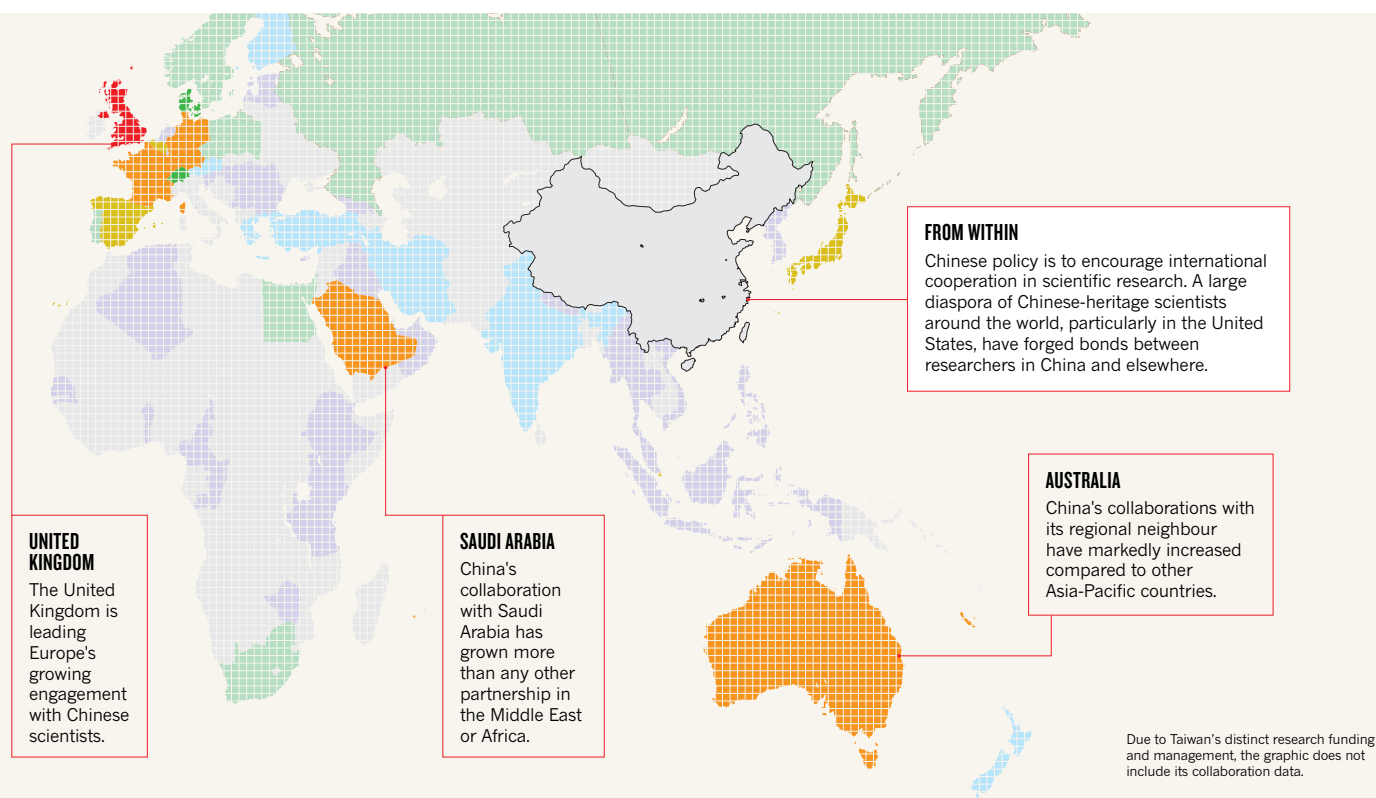
## DISCIPLINARY STRENGTHS

China's booming scientific output is concentrated around specific subject areas, a trend that has continued since 2012. Chemistry and physical sciences clearly dominate the country's total publishing output in the Nature Index (see 'Chemistry champs'). The WFC figure for chemistry in 2014 was 3,783, accounting for 61% of the country's total WFC, while physical sciences made up 30% of China's publishing output in the index. By comparison, distribution of the WFC in other subject areas are represented more proportionally in other top contributing countries, such as the United States, Germany and the United Kingdom.

The Chinese Academy of Sciences (CAS) is the top institutional producer in chemistry WFC, both in China and around the globe (see 'China's superstar'). The Institute of Chemistry (ICCAS) is the top contributing CAS institute by WFC. Its research strengths lie in molecular and nanosciences, organic and polymeric materials, chemical biology, as well as energy and green chemistry.

These cutting-edge areas of chemistry tend to have an applied aspect and are essential for industrial innovation. For instance, an ICCAS researchers' study on the assembly mechanism of organic composite materials strongly contributed to the development of flexible





photonics and the realization of nanophotonic circuits for next-generation optical information processing.

The Chinese government has a crucial role directing the country's science and technology development. "The government's emphasis on the commercialization of high technology and the capacity of scientific research to drive industrial productivity possibly explains the strong focus on chemistry, particularly the subfields that are easily translated into commercial production," says Liu.

In line with the demands of the national development strategy, China is also making efforts to innovate in relatively newer fields in life sciences and environmental sciences such as energy, water resources, agriculture, environmental protection, and human health, which were identified as research priorities in 2006 in China's 15-year plan.

*"The nature of the scientific revolution is long-term. China will lead more and more programmes."*

"Strong demand for new energy and the need to reduce pollutants emission in energy consumption will drive China's growth in environmental sciences," anticipates Liu, whose background is in this field.

Life sciences are also expected to make great advances in the near future. Between 2012 and 2014, China's output in this area grew by 30%. Fields such as genomics and protein sciences, stem cell and cloning technology, and gene

therapy have already experienced significant progress. "[China is] set to become the global powerhouse of gene and protein research, leading this exciting field in life sciences and making grand discoveries with profound impacts," He explains.

### INCREASED INTERNATIONAL COLLABORATION

Collaboration is an increasingly significant aspect of modern science and China's collaboration scores in the index reflect this trend. The recent Nature Index 2015 Collaborations supplement revealed that China's international partnerships are soaring, with its collaboration score rising 31% from 2012 to 2014. Collaboration score is the sum of the fractional counts

(FC) for each of China's bilateral partnerships. Almost half of China's international collaboration score in 2014 came from partnerships with the United States. Correspondingly, China has become the United States' largest collaborator, surpassing Germany in 2014. Other important international collaborators for China are other top contributing countries to the Nature Index, such as the United Kingdom and Japan (see 'Collaboration hotspots').

With the return of many Chinese scientists who have trained abroad, international collaborations are often based on personal ties. This has raised questions about China's role in these collaborations, some suggesting it is merely providing cheap labour working on ideas at the behest of former supervisors.

But in the past five years, Reiner says there has been a shift in these partnerships as Chinese scientists play a more significant role in the research and Chinese institutes contribute a greater proportion of the funding. In some international collaborations, for instance the Human Liver Proteome Project (HLPP) led by He, China is already setting scientific objectives and making important theoretical and technological innovations. "The nature of the scientific revolution is long-term," Reiner says. "As the output of scientific research from China is growing, I have no doubt that China will be leading more and more major international programmes, providing more important and, in some cases, critical contributions to international collaboration proportional to its input." ■



Winner of the Nobel Prize for Medicine, Tu Youyou.

HAN HADAN/CHINAPHOTO/CHINA FOTOPRESS VIA GETTY



A lion guards Beijing's Forbidden City as a symbol of strength. China's capital is itself a stronghold of scientific achievement, along with Shanghai and Nanjing.

# THREE GIANTS TIGHTEN THEIR GRIP

*The benefits of economics and history converge with the demands of population growth and sustainability issues in China's most productive research and technology centres.*

BY PENG TIAN

It will come as no surprise that the top performing Chinese cities in the Nature Index are Beijing, Shanghai and Nanjing. All three are significant players economically and politically, Beijing and Shanghai particularly.

At all levels of Chinese government, officials see innovation through science and technology as critical for the nation to achieve continued economic growth on a more environmentally sustainable path. The central government has invested heavily in science and technology to improve productivity and upgrade some industries, such as the manufacture of advanced high-speed trains.

This, in turn, has reinforced the status of Beijing, Shanghai and Nanjing, which represent the Chinese cities with the top 2014 WFCs. (see 'China's top 10').

The local governments of advanced provinces also push technological innovation to achieve economic ascendancy over other cities. Market forces are playing an increasingly significant role in Beijing, Shanghai and Nanjing, as both international and domestic commercial enterprises work with universities to develop next-generation technologies.

***"A long period of economic growth has been mirrored by committed investment."***

Crucially, China's research system was significantly rebuilt after the turmoil of the Cultural Revolution — the social and political movement that began in 1966.

A long period of economic growth has been mirrored by committed investment in science and technology. Besides the National Natural Science Foundation of China (NSFC), there are several ongoing programmes that promote basic research and technological innovations in

universities and institutes. These have included the 863 and 973 Programmes under the Ministry of Science and Technology (MOST), and Projects 211 and 985 under the Ministry of Education. Through these initiatives elite universities and institutes in Beijing, Shanghai and Nanjing have received enormous funding support to build advanced research facilities and improve research quality.

## BEIJING

Direction shows position change 2013–2014

**WFC rank China: 1**

**AC: 5,163**

Beijing, the nation's capital, has been the centre of power for China for millennia, and a research and industry stronghold since the foundation of the People's Republic of China in 1949.

Beijing benefits most from the systems of resource allocation established after 1949 when the communist government restructured and relocated the country's main education and research centres.

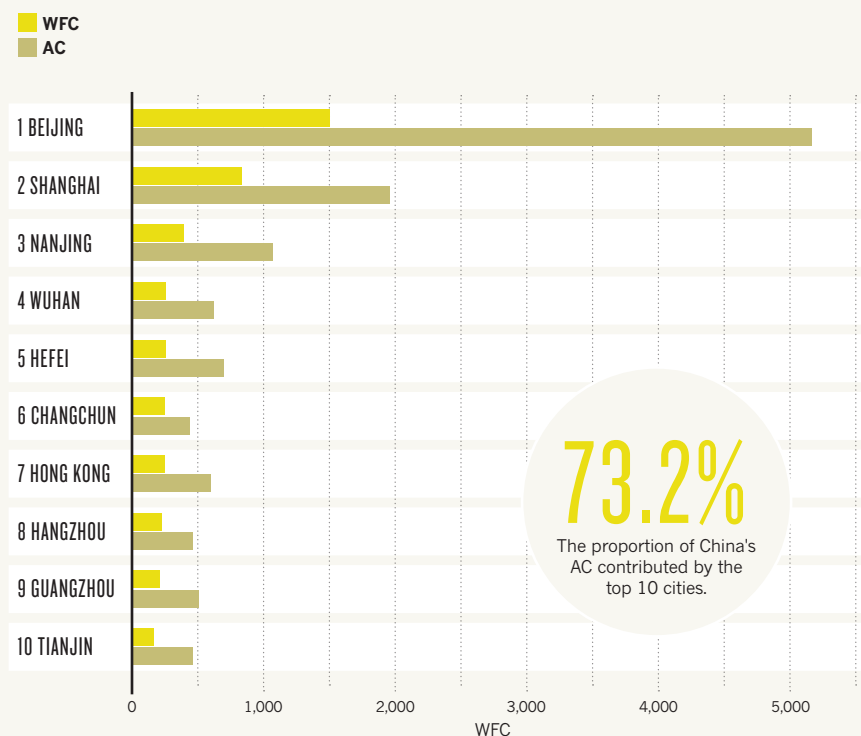
It has the highest number of institutions in the Nature Index — 131 in 2014 — of the three





## CHINA'S TOP 10

The country's most productive cities in the Nature Index by WFC in 2014.

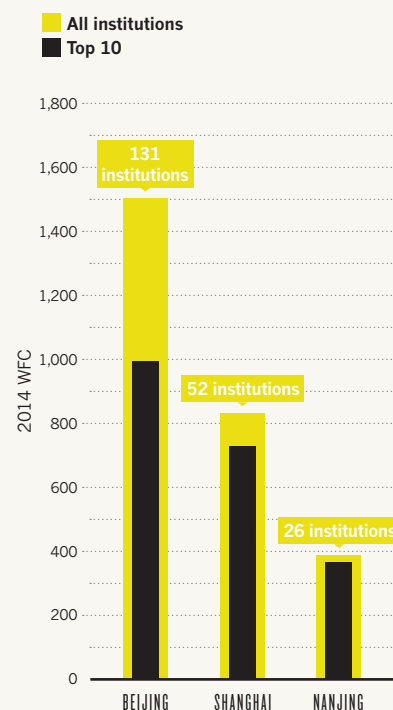


**70.4%**

The proportion of China's WFC contributed by the top 10 cities.

## CITIES OF INFLUENCE

Beijing hosts the most institutions, while Shanghai and Nanjing's WFC comes largely from their top 10 performers.



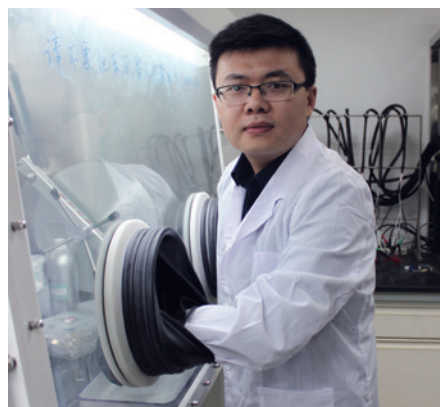
cities and is particularly strong in chemistry and physical sciences. This contributes greatly to Beijing's overall output in the Nature Index. Beijing has also inherited some of the most prestigious universities established prior to 1949. These include China's leading Peking University (PKU) and Tsinghua University, which receive resources from central and local governments. Several significant institutes of the behemoth Chinese Academy of Sciences (CAS), such as the Institute of Chemistry and Institute of Physics, are also located in Beijing. Between them, Beijing's top 10 institutions accounted for around 60% of the city's overall 2014 WFC (see 'Cities of influence').

**"PKU's special position in China has played a key role in helping us get support from the central government."**

The vast resources allocated to these universities have paved the way for some groundbreaking research, including an important development in quantum computing by Duan Luming from Tsinghua University. In a paper

published in *Nature*, he and colleagues described experiments that bring robust quantum computation at room temperature closer to reality.

Generous funding has also helped physicist Peng Lianmao at Peking University, who is developing technology that can build carbon nanotube semiconductor devices and integrated circuits. In a paper published in *Applied Physics Letters*, his team detailed the construction of



Wang Yonggang focuses on new lithium batteries.

high-performance carbon nanotube transistors and integrated circuits, which represents the future of computer processors. This work received funds from the MOST, NSFC, and the Beijing Municipal Science and Technology Commission.

"The research work of constructing nano devices and integrated circuits needs huge funding," Peng says. Since 2011, his lab has received about 70 million RMB from Project 973. "PKU's special position in China has played a key role in helping us to get support from the central government."

Now, he says, the lab is the only one in the world that can construct 10-nm carbon nanotube complementary metal-oxide-semiconductor (CMOS) integrated circuits. In 2012, the application potential of this attracted local government money from the Beijing Municipal Science and Technology Commission. "Beijing Municipality is helping us to make some long-term development strategies for carbon-based integrated circuits," Peng explains. He is optimistic that a clear strategy and further funding will come through at the national level to take the technology even further.

WANG YONGGANG



The manufacture of advanced high-speed trains has attracted significant government funding.

## SHANGHAI

WFC rank China: 2  
AC: 1,955

In an idyllic location on the estuary of the great Yangtze River in the centre of East China's Yangtze River Delta, Shanghai has become a world-renowned port and global financial hub since it opened to international trade in the 1840s. Shanghai, which has been one of the world's major manufacturing centres for more than two decades, receives generous research resources from the regional economy.

With a 2014 population of more than 24 million, Shanghai is now the largest and most populous city in China. It has the prestigious Fudan University as well as many research institutes of the CAS and other universities built after 1949.

Shanghai has less than half the total number of institutions of Beijing in the index — 52 in 2014 — but its top 10 produce a similar output to the capital's top 10. Shanghai's strength lies in chemistry (see 'Shanghai's best game'). More

than 60% of the city's output in the Nature Index is in chemistry, and Fudan University and the CAS Shanghai Institute of Organic Chemistry are the biggest contributors.

"Besides the promotion from the state and Shanghai, international corporations, giant state-owned enterprises, and private enterprises also actively collaborate with Shanghai's universities and research institutes to develop new technologies," says Wang Yonggang, a materials chemist and associate professor in the Department of Chemistry at Fudan University. Between 20 and 30% of his lab's funding comes from such collaborations.

Wang's research focuses on new types of lithium battery. He's already had some promising results published in *Chemical Communications*, which show that these devices have the potential to be next-generation batteries for everything from electric cars to smartphones.

Wang is a partner of the Collaborative Innovation Center of Chemistry for Energy Materials (iChem), a project of Plan 2011. "The centre can coordinate experts from different research fields, make communication more efficient, and has the capacity for commercializing the achievements," Wang explains.

## NANJING

WFC rank China: 3  
AC: 1,064

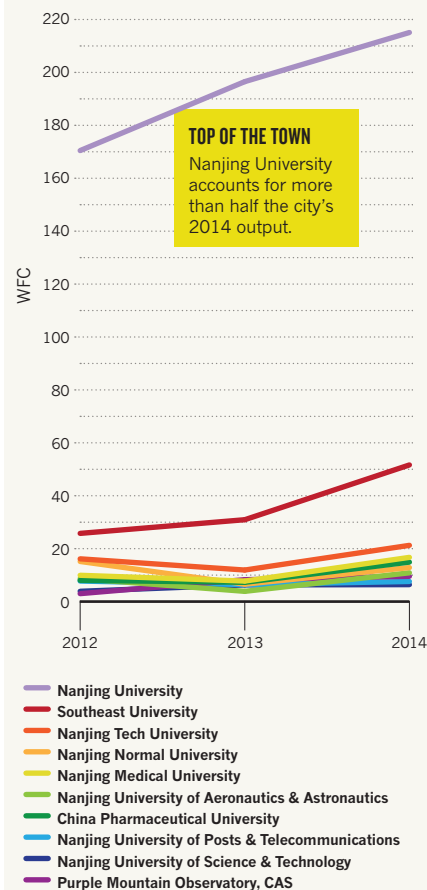
As the capital of the wealthy eastern coastal province of Jiangsu, Nanjing is located in a region rich in economic and technological activity.

Nanjing means 'southern capital' in Chinese — an indication of its status — and is the second largest city after Shanghai in the prosperous Yangtze River Delta. In recent years, Nanjing's largest growth across all the subject areas has, as with Shanghai, been in chemistry. Also like Shanghai, almost 60% of its output in the Nature Index is in chemistry. In an effort to differentiate itself from its regional counterpart, grants to promote materials science and astrophysics have been offered to research groups in Nanjing.

Nanjing University is the main source of the city's scientific discovery and technological innovations and the city's main contributor

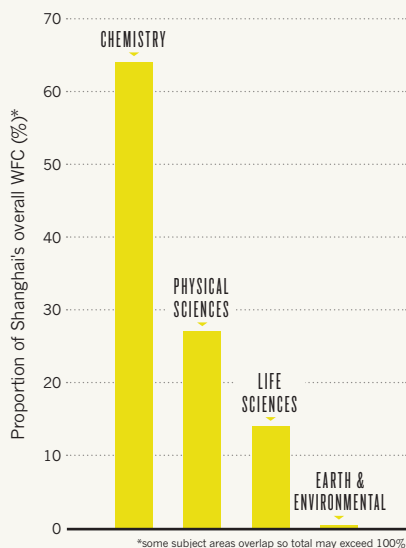
## LEADING THE PACK

Nanjing's overall growth is driven by the performance of a single institution.



## SHANGHAI'S BEST GAME

Relative subject area strength as a proportion of overall WFC in 2014.



to the Nature Index, accounting for more than half of Nanjing's overall 2014 output (see 'Leading the pack'). The Collaborative Innovation Center of Advanced Microstructures (CICAM), which is part of Plan 2011, generates most of Nanjing's research on artificial microstructure materials. As well as basic research, this centre also tries to translate research findings into practical applications to meet the technological needs of the delta's industries.

CICAM acoustic physicist Bin Liang, a professor at Nanjing University, has designed and experimentally realized a new acoustic absorption material that may be used for noise reduction and to make echo-free underwater materials. His research was recently published in *Applied Physics Letters*.

Not so long ago the acoustic qualities of this new material were thought to be impossible to achieve, Bin says. "We always think about trying to extend the limits of knowledge," he explains. "Now, we want to translate the breakthroughs into applications." ■

*"We always try to extend the limits of knowledge. Now we want to translate the breakthroughs."*





PHILIPPE LELANNE/GETTY

Chengdu's sparkling and fast-growing skyline is a shining testament to its transformation into a high-tech hub, which is home to almost 100 cutting-edge labs.

# AT THE VERY HEART OF PROGRESS

*The ambition driving China's astonishing progress in the output of high-quality science is particularly strong in some cities, whose growth far outstrips expectation.*

BY SARAH O'MEARA

When neuroscientist Anna Wang Roe was looking two years ago for somewhere to set up a new state-of-the-art interdisciplinary neuroscience and technology institute, her search ended in the far eastern city of Hangzhou. "I looked at many top universities. But then I went to Zhejiang University [in Hangzhou] and it really stood out, even over some higher ranked Chinese institutions," she recalls.

Between 2012 and 2014, China's Nature Index WFC rose by 37%. But several cities grew at an even faster rate — Hangzhou, Xi'an and Chengdu being some of the standout

examples (see 'Stellar performers').

The overall growth in publication output from these cities has largely been in chemistry (see 'Subject specialities'). Yet researchers working in these universities point to many factors, beyond expertise in a single discipline, for their success. Wang Roe, for instance, was so impressed by Hangzhou's atmosphere of energy, passion and collaboration that she approached the city's Zhejiang University, China's fifth largest Nature Index contributor in 2014, with her institute proposal. "At Zhejiang they have created a highly motivated research environment where people can share and explore freely. This doesn't always happen in China," she notes.

## XI'AN

WFC rank China: 12

AC: 309

Situated in China's northwestern Shanxi province, Xi'an shows the largest relative increase in WFC of the three cities, increasing by 142% from 2012 to 2014 (see 'Stellar performers'). WFC is a metric that apportions credit for each

article according to the affiliations of the contributing authors.

Xi'an has a history of at least 3,000 years, more than a thousand of those as the capital of ancient Chinese dynasties. Since the early 1990s, when it emerged as the lead city of China's Western Development programme, Xi'an has promoted the value of tech-driven industry and established one of the earliest national high-tech development zones. The city is also home to a sophisticated national aviation base that was integral in manufacturing *Shenzhou 6*, the craft that carried China's second manned space flight in 2005.

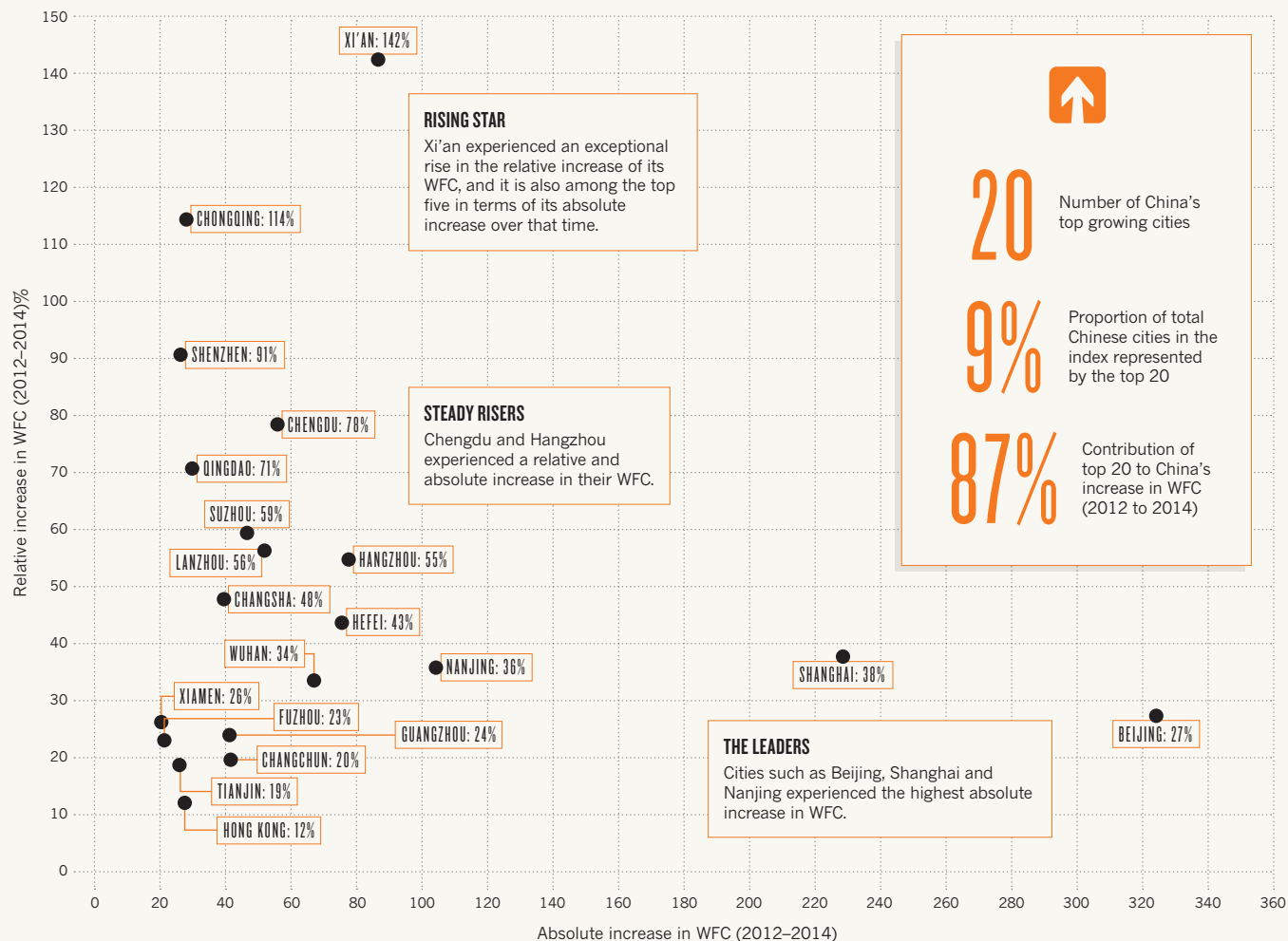
Between 2012 and 2014, the fast rise in the number of publications from Xi'an institutions featured in the Nature Index's 68 high-impact journals was led by Xi'an Jiaotong University. The largest subject increases were in chemistry and physical sciences.

Shan Zhiwei is deputy director of the university's State Key Laboratory for Mechanical Behavior of Materials, which contributed to almost 25% of Xi'an Jiaotong University's articles in the Nature Index in the three-year growth spurt. He believes the university's success is



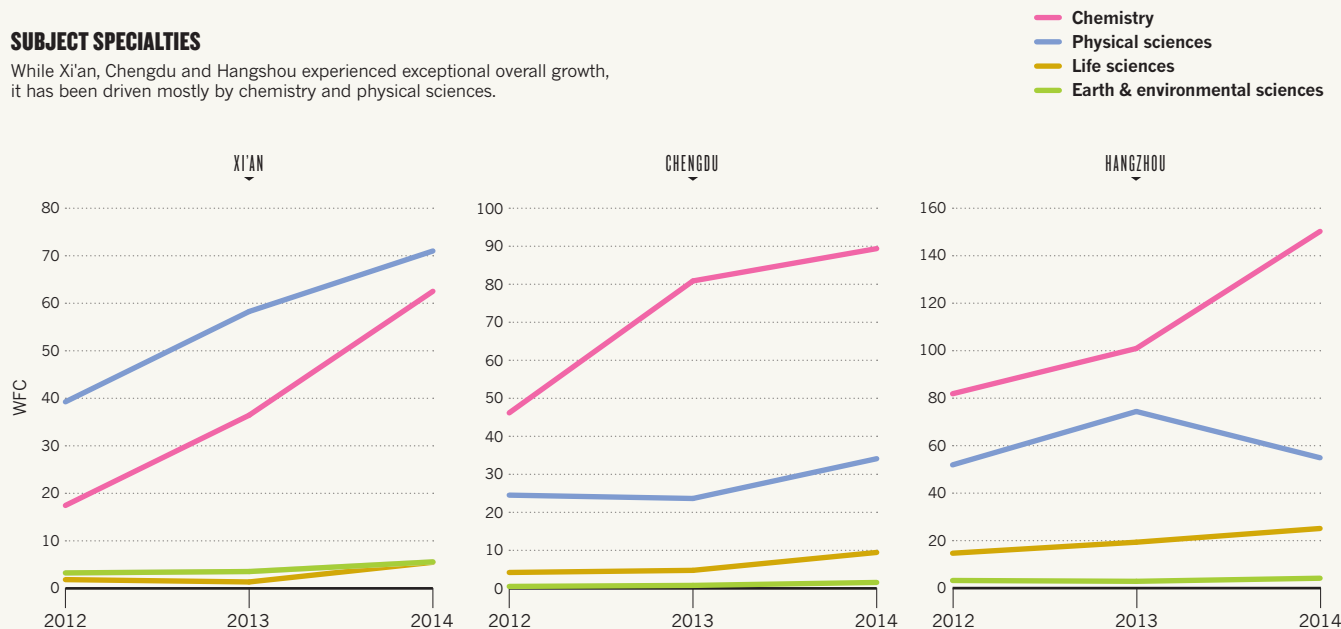
## STELLAR PERFORMERS

The proportional increase of China's top 20 growing cities compared to their overall increase in WFC over three consecutive years between 2012 and 2014.



## SUBJECT SPECIALTIES

While Xi'an, Chengdu and Hangzhou experienced exceptional overall growth, it has been driven mostly by chemistry and physical sciences.





driven by government strategies to attract international science talent to China, combined with innovative recruitment policies. “Xi’an Jiaotong

**“An emphasis on talent recruitment has been accompanied by a willingness to embrace new ways of working.”**

has developed a series of local policies to attract talents, including offering attractive salaries, strong financial support, and an open and approachable management style,” explains Shan, who is also a director of the Center for Advancing Materials Performance

from the Nanoscale (CAMP-nano),

He adds that in the past five years, Xi’an Jiaotong has hired many foreign-born experts and enticed hundreds of Chinese-born professionals working abroad to return, bringing their education and experience. The recruitment drive has been accompanied by a willingness to embrace new approaches to work. “We drew on the strength of the new people,” Shan says. “They have introduced new methods of training students, managing team members and handling instruments. In addition, the recruits have strong financial support. This combination has led to high efficiency of work and yielded good research outputs.”

## CHENGDU

WFC rank China: 13

AC: 287

Science and technology-driven development has transformed Chengdu into one of the world’s fastest growing cities and led to a surge in high-quality research output. The WFC of Chengdu in the Nature Index increased by almost 80% between 2012 and 2014 (see ‘Stellar performers’).

The city has allocated enormous resources to create an environment where innovation thrives, starting in the laboratory. Chengdu now has 10 national key laboratories funded by China’s central government, 30 labs established by branches of local government and 53 universities.

Between 2012 and 2014, the Nature Index contribution of the city’s Sichuan University soared, particularly in chemistry and related disciplines. “Chengdu has become a hotbed for academic achievement,” says Wei Yuquan, vice president of Sichuan University and director of the National Key Laboratory of Biotherapy, which contributed to articles in the Nature Index between 2012 and 2014. “Our multi-disciplinary research centre has established an integrated technology chain for the discovery and development of innovative drug candidates in a single institute,” he says.

Chengdu is the capital of the landlocked Sichuan province and has been growing rapidly since 2000, when China’s central government began pouring money into poorer interior



Xi’an’s sophisticated aviation base was central to the manufacture of the spacecraft *Shenzhou 6*.

cities. Official policy has focused on turning the city into a high-tech hub, a role Chengdu was well positioned to fill. “Local government is establishing Chengdu as an area of innovation, and has set up many foundations to support research projects including basic research and to recruit talented researchers,” Wei says.

Among them is Gong Qiyong, deputy dean of West China Medical School, Sichuan University in Chengdu. Ten years ago he resigned from a faculty position at the University of Liverpool in England to become the director of West China Hospital’s Huaxi MR Research Center. “Now my group has been internationally recognized as one of the leading teams in psychiatric imaging,” he says. “Recently we obtained a huge grant from the government to build the National Research Center of Translational Medicine.”

## HANGZHOU

WFC rank China: 8

AC: 458

Hangzhou is one of China’s quintessential historic cities, but its rapidly increasing rate of high-impact research suggests its best years are yet to come. Hangzhou, where the Nature Index WFC jumped by 55% from 2012 to 2014, has the largest absolute WFC of the three highest performing cities (see ‘Stellar performers’).

During the past decade, Hangzhou has become a hub for science-driven innovation from laboratory research to tech start-ups. It’s home to the Alibaba Group, China’s leading

**“There’s huge energy in Hangzhou. They go for it with full force.”**

e-commerce service provider with an estimated value of US\$231 billion. The local government has created a network of opportunities to encourage similar success stories, including grants to

promote technology transfer from the lab into business, and start-up funds for academics setting up companies at incubator sites.

Wang Yong is a professor at Zhejiang University’s School of Materials Science and Engineering, which hosts a nationally-funded laboratory for silicon materials, and contributed to papers from the university in the Nature Index between 2012 and 2014. Wang is researching the catalytic mechanism of nanocrystals at nanoscale in order to develop high performance catalysts for future industry use. “The local government is rich and invests a large amount of money in universities,” he explains. “There is good start-up funding available, financial incentives for high-quality work, state-of-the-art facilities, and the opportunity to innovate and collaborate with industry.”

Hangzhou also benefits from leaders whose clear vision was a big factor behind Wang Roe’s decision two years ago to ask Zhejiang University for a US\$25 million grant to build her Interdisciplinary Institute of Neuroscience and Technology, of which she is now the director.

“The administration tells me the grant was the largest investment of any university in China on a single project,” Wang Roe says.

The five-storey building with 20 labs and a large primate facility officially opened in October 2015. Academics came from all over the world for the opening ceremony and conference. “[They] were very impressed [and] amazed that this could be achieved in such a short time,” Wang Roe says. “I looked at many high-level institutions in China for this project, but Zhejiang had it all; excellent engineering, optics, materials science, information sciences, neuroscience and medicine, and a collaborative environment.

“There’s huge energy in Hangzhou. Once the administration decides on something, they go for it with full force with the long-term in mind. It makes you think differently, it really does.” ■





The flow of the Shenzhen River mirrors the extraordinary dynamism of the city it passes, home to innovative manufacturers and a commercial transformation.

# THE CHANGING FACE OF INDUSTRY

*Amid fierce international competitiveness, governments at all levels are responding by orchestrating collaborations between industry and academic institutions.*

BY DAVID CYRANOSKI

Like many nations, China is hotly pursuing innovation and the economic benefit of bringing ideas to the market. But China's drive to embrace an innovation-based economy in favour of its reliance on manufacturing is daunting. Most big companies are state owned and traditionally averse to funding research. Despite a dramatic increase in basic research output over the past two decades, only a small percentage is converted to industrial application.

Yet as Shenzhen, Beijing and Wuhan show, an industrial base built on cutting-edge science has matured quickly in some regions, often with local policy support. These cities host many

of the Chinese corporations with the highest growth in research output published in the Nature Index's 68 high-impact journals between 2012 and 2014 (see 'Industry champions'). With a swag of intellectual property and revenue as proof, these regions are leading China in its quest for transformation.

## SHENZHEN

WFC rank China: 21

AC: 165

Shenzhen, in the country's southeast, has had the most marked transformation to a research-based industry hub of any city in China, and probably the world. Just 35 years ago it was a fishing village; now it's a thriving metropolis that links Hong Kong with China's mainland. Companies based there account for almost half of the country's international patent filings.

Most of these filings are in telecommunications and electronics, with Huawei and ZTE leading the way. These two ICT multinationals, along with Shenzhen-based rechargeable battery-maker BYD Co., Ltd, boast China's three biggest patent portfolios.

Shenzhen is also home to the Kuang-chi Institute of Advanced Technology, a manufacturer of radar absorbent materials used in stealth technology. It was founded by a group of Chinese scientists returning from stints working in the United States.

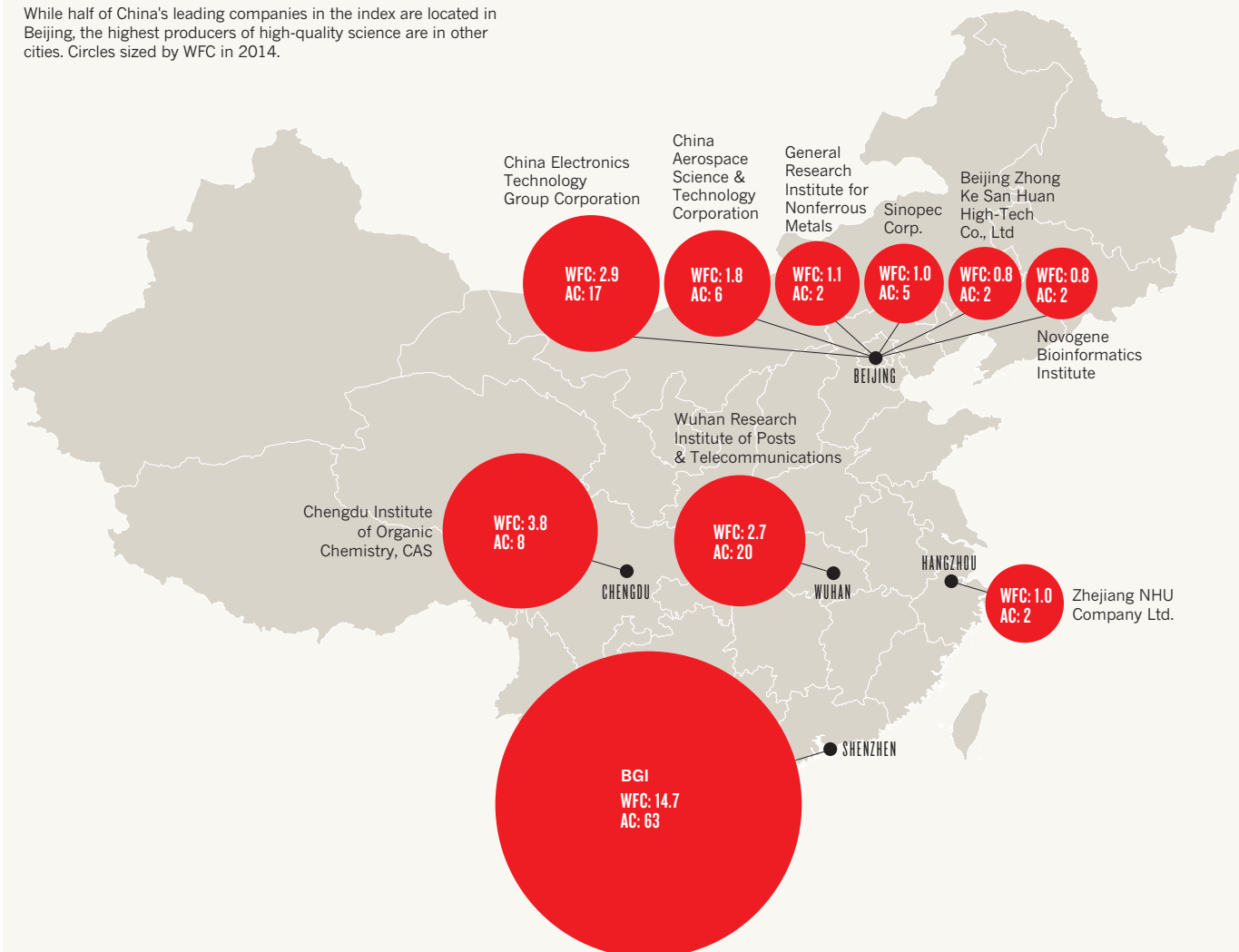
Casting a broader net, the Shenzhen Institutes of Advanced Technology, one of the most industrially prolific units of the Chinese Academy of Sciences (CAS), has established collaborations with more than 150 companies during its 10-year history.

The genomics sequencing powerhouse BGI, in particular, has successfully melded basic research in Shenzhen with commercial operations. It's one of China's biggest contributors to high-impact scientific publications, holds some 400 patents and has another 300 pending. More than half are related to genes, especially in the areas of agriculture and rare human diseases. Another third relate to sequencing technology, and the rest are special applications. BGI also holds the crown for the Chinese company with the highest 2014 WFC. Indeed, it is in the world's top 20 corporate contributors to the Nature Index.



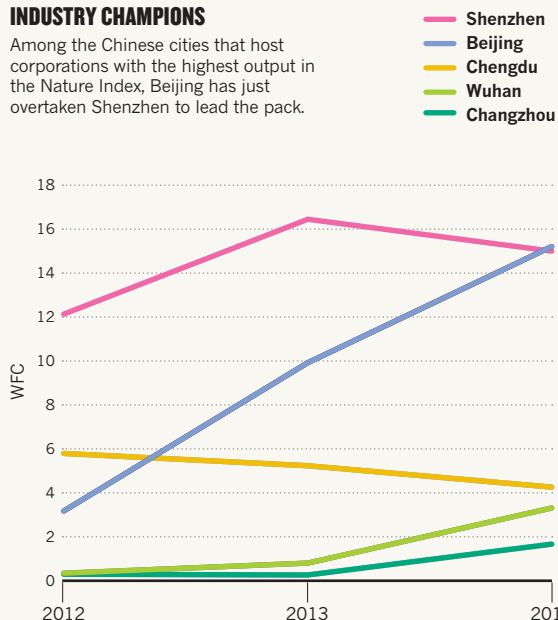
## CHINA'S TOP 10

While half of China's leading companies in the index are located in Beijing, the highest producers of high-quality science are in other cities. Circles sized by WFC in 2014.



## INDUSTRY CHAMPIONS

Among the Chinese cities that host corporations with the highest output in the Nature Index, Beijing has just overtaken Shenzhen to lead the pack.

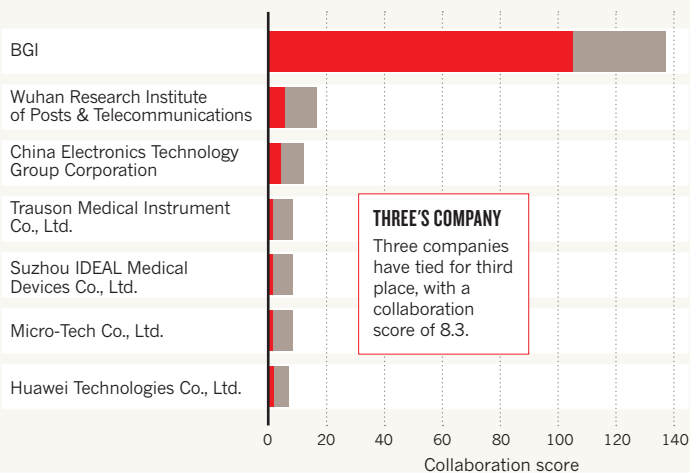


## CORPORATE GIANTS

China's top 5 collaborating companies in 2014 by collaboration score.

■ Corporate contribution  
■ Partners' contribution

Collaboration score = the sum of the FC for each company's bilateral partnerships.







Funding support for greener production improves conditions for car and battery manufacturer, BYD.

The success of BGI and other Shenzhen companies, says executive director of BGI Research, Xu Xun, has been bolstered by the city's history as China's first special economic zone in 1980, an initiative that made it easier to start a company and interact with foreign companies. "In other cities, many companies are founded by or supported by the government. Patents are not important for them," says Xu. "Here, there are a lot of private companies and they need intellectual property so we put a lot of effort into it."

BGI is presently celebrating the first harvest of a drought-tolerant millet strain that was bred on the strength of discoveries from a sequencing project at the company. BGI hopes this new millet will find a large market in a China where there are increasing concerns about water resources. BGI will also expand into the clinical sequencing market with a new line of desktop sequencers, trying to take advantage of China's move towards personalized medicine.

The Shenzhen government uses the number of patents as one measure of a company's value to the city, with companies deemed significant enjoying benefits such as special fast-track approval processes. It also gives annual awards for the most impressive innovations. "We get pressure to have good intellectual property both from the government and from our needs as a private company," says Xu. "Shenzhen always finds ways to support innovation."

## BEIJING

**WFC rank in China: 1**  
**AC: 5,163**

While Shenzhen might have flexible rules and an entrepreneurial environment, Beijing has its own advantages to make it a bustling and innovative industrial centre, says Jin Qinxian, head of the technology transfer office at Tsinghua University. The city's numerous universities — most notably Tsinghua and Peking University — and a slew of institutes either independent or affiliated with the CAS have provided fertile ground for technology transfer. "The history,

the culture, the number of institutes and universities make Beijing very powerful," says Jin. Beijing also has a number of highly talented, internationally renowned researchers who have returned from working or studying abroad.

Among their number is Wang Xiaodong, a former Howard Hughes researcher at the University of Texas Southwestern, in the United States, who designed and now directs the National Institute of Biological Sciences in Beijing. He also founded BeiGene, a company

*"There is pressure for good intellectual property from the government and from our company needs."*

with several large and small molecule cancer treatments in development and which received US\$97 million in financing this spring. It has already partnered with Merck Serono on two drugs.

Beijing also has a sequencing company, Novogene, run by an ex-BGI employee, and now competing with BGI. Novogene rounds out the top 10 Beijing companies with the highest 2014 WFC (see 'China's top 10').

Tsinghua, which ranks first for research funding in China, has a particularly rich field of scientists active in commercialization, transferring technology that has helped China become a leader in carbon nanotubes and high-speed computing. It has also contributed to a broad range of biomedical breakthroughs including cancer biomarkers and medical devices such as pacemakers.

Many of these are interdisciplinary projects. A team led by chemical engineer Dehua Liu, for example, designed a new enzymatic process for converting renewable oils and fats to biodiesel. A bioenergy production plant is now pumping out 20,000 tons of biodiesel per year — a figure that will jump five-fold next year — and the technology has been transferred to companies in several countries, including Germany and Brazil.

In 2014 alone, Tsinghua had 2,010 domestic patent applications (1,360 of which have been

accepted) and 264 more in the United States. It received 150 million RMB from 61 transfer or licensing agreements.

Specific local programmes help, says Jin. Beijing's municipal government provides start-up funds in exchange for shares in the companies. It also nurtures companies with initiatives such as making the first purchase order for products, before they are even proven, from Beijing firms.

But what has really pushed the city's industrial blossoming are the rich human resources, concentrated especially in the massive Zhongguancun industrial zone and technology hub that neighbours both Tsinghua and Peking universities. "Companies come and they get access to students, to laboratories, to professors," says Jin.

## WUHAN

**WFC rank China: 4**  
**AC: 619**

Wuhan is one of the fastest growing cities in China in terms of the scientific output of its corporations, and it is quick to capitalize (see 'Industry champions'). The Huazhong University of Science and Technology (HUST), for example, has a long list of successful spin-off companies based on optical fibre development for high intensity laser and 3D printing. With the success of research intensive companies such as Huagong Tech, which produces lasers, holograms, and optical communication devices, and Guide Infrared, which manufactures a cutting-edge night vision camera, Wuhan has taken a place at the forefront of China's optoelectronics and telecommunications boom.

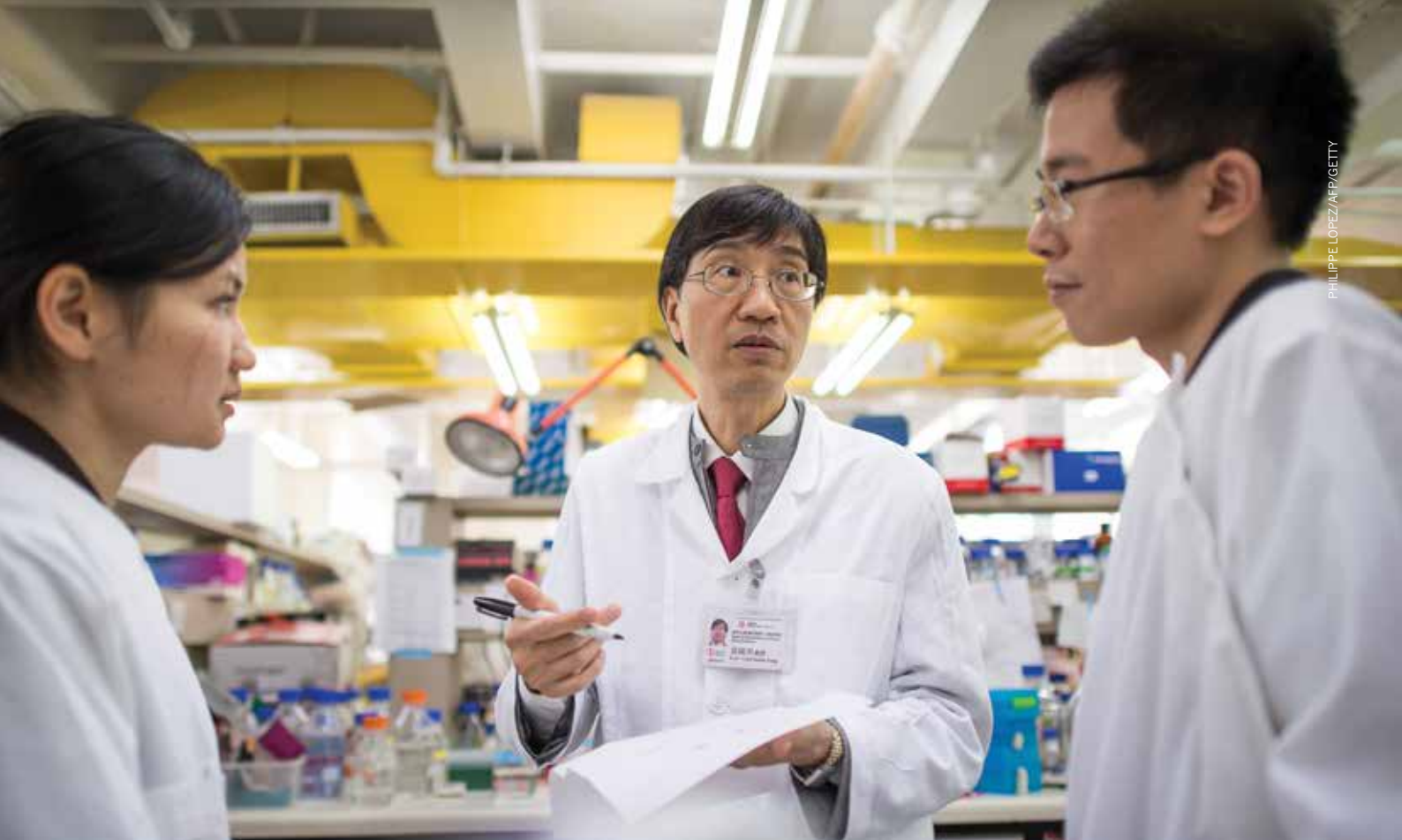
Tang Jiang, a thin-film photovoltaics researcher at Wuhan National Laboratory for Optoelectronics, says the presence of many leading universities in Wuhan, such as HUST, which ranked 19 in engineering in the *US News* global survey of universities, has played a major role in gearing up this industrial output.

Tang's institute has several examples of technology transfer, including blue photoluminescent materials for organic light-emitting diodes (OLEDs), a UV lighting diode used for solidification in printing and a 'micro-optical tomography system' for brain imaging.

Local government policies have promoted technology transfer from universities, establishing a requirement that a research team receives at least a 70% share of the technology transfer profit. "These policies significantly encourage professors in universities to focus on application orientated research and the consequent technology transfer," Tang explains.

Tang has yet to commercialize his own research, infrared photodetection and the creation of new materials based on the promising thermoelectric antimony selenide. This work could lead to non-toxic and cheap next-generation flexible solar cells. But once his devices reach an energy conversion efficiency threshold of 10%, from the current best of 5.6%, he will reach out to industry. ■





PHILIPPE LOPEZ/AFP/GETTY

Hong Kong-based Yuen Kwok-Yung, who identified the cause of SARS, says collaboration between scientists of different perspectives leads to novel breakthroughs.

# ALLIANCES FOR SCIENTIFIC SUCCESS

*The diverse histories of China's cities strongly influence their collaboration patterns.*

BY HEPENG JIA

**X**u Aimin knew that collaborating represented his best chance for success in his quest at Hong Kong University (HKU) to identify the relationship between obesity and the glucose-regulating hormone adiponectin.

Connecting with colleagues in Saudi Arabia and Korea, he formed a team that revealed something unexpected. While obesity is a consequence of metabolic dysfunction, it is also exacerbated by it, because the expression of adiponectin is reduced. Altering this activity may represent a new strategy for the treatment of obesity-related disorders, their study, published in *Nature Communications*, suggests.

The findings came about through the team's combined knowledge and resource base, drawing on the Korean researchers' physiology expertise, the HKU lab's excellence in biology and unique animal models, and the Saudi Arabian lab's clinical samples.

"Hong Kong is a relatively small place," says Xu, a professor in the university's department of medicine. "Collaboration with [Chinese] mainland and overseas institutes is the best way to maximize the economic and societal impacts of our research."

Xu's first-hand experience echoes what large scale studies show about science in the twenty-first century: research resulting from collaborations is more frequently cited, especially papers with international co-authors.

In the Nature Index, three Chinese cities stand out for their collaborative orientation: Hong Kong, Hefei and Tianjin.

The focuses of their collaborations differ, all bring great reward. While Hong Kong and Hefei institutions have formed a record number of partnerships with their international peers (see 'Hong Kong's hotspots'), Tianjin scientists have focused on forging local links (see 'Close ties').



IMAGERITE/ALAMY

A sculptural sundial at HKUST represents early human invention, an inspiration for scientists.

Exploring the roots of these patterns reveals the importance of history in shaping regional strengths.

As Wu Yishan, vice president of the Chinese Academy of Science and Technology for Development, puts it: "With the joint forces of historical tradition, research capacity, local



policies and personal links, Chinese regions have formed different preferences in research collaborations.”

## HONG KONG

WFC rank China: 7  
AC: 600

The index shows that Hong Kong's collaborations are firmly entrenched with, but certainly not limited to, mainland China, the United States and Europe. For example, the two leading international collaborators for the Hong Kong University of Science and Technology (HKUST) are the National University of Singapore and Singapore's Agency for Science, Technology and Research. HKU's most frequent overseas collaborator is Taiwan's National Tsing Hua University (see 'Hong Kong's hotspots').

The collaborative atmosphere in Hong Kong appears to be fostered by funding policies. The region's local agencies support a large number of collaborative and joint funding schemes with other bodies at home and overseas.

An example is the HKU-Pasteur Research Centre, which was established to tackle emerging infectious diseases in China and elsewhere in Asia. Yuen Kwok-Yung, an HKU microbiologist who is renowned for tracing the human SARS (severe acute respiratory syndrome)

coronavirus to Chinese horseshoe bats, was appointed its first co-director. Yuen then went on to help launch the HKU AIDS Institute in collaboration with the Aaron Diamond AIDS Research Center, an affiliate of Rockefeller University in New York. “Scientists from different cultures and ethnicities have very different and novel perspectives for looking at a scientific question and provide varied approaches to find the solution,” Yuen says.

## HEFEI

WFC rank China: 5  
AC: 696

Like Hong Kong, Hefei, capital of the Chinese mainland hinterland province of Anhui, has a limited number of research institutions in the index. One of these, the University of Science and Technology of China (USTC), is the driving force behind the city's collaborations. USTC's WFC between 2012 and 2014 was 517, more than six times that of the second player, the Hefei Institutes of Physical Science, which is affiliated with the Chinese Academy of Sciences (CAS).

History and tradition help explain USTC's unique role. Founded by CAS, the university moved from Beijing to Hefei in 1969 during the Cultural Revolution. Following the opening



Hefei's USTC campus provides a highly supportive environment for collaborative research.

## HONG KONG'S HOTSPOTS

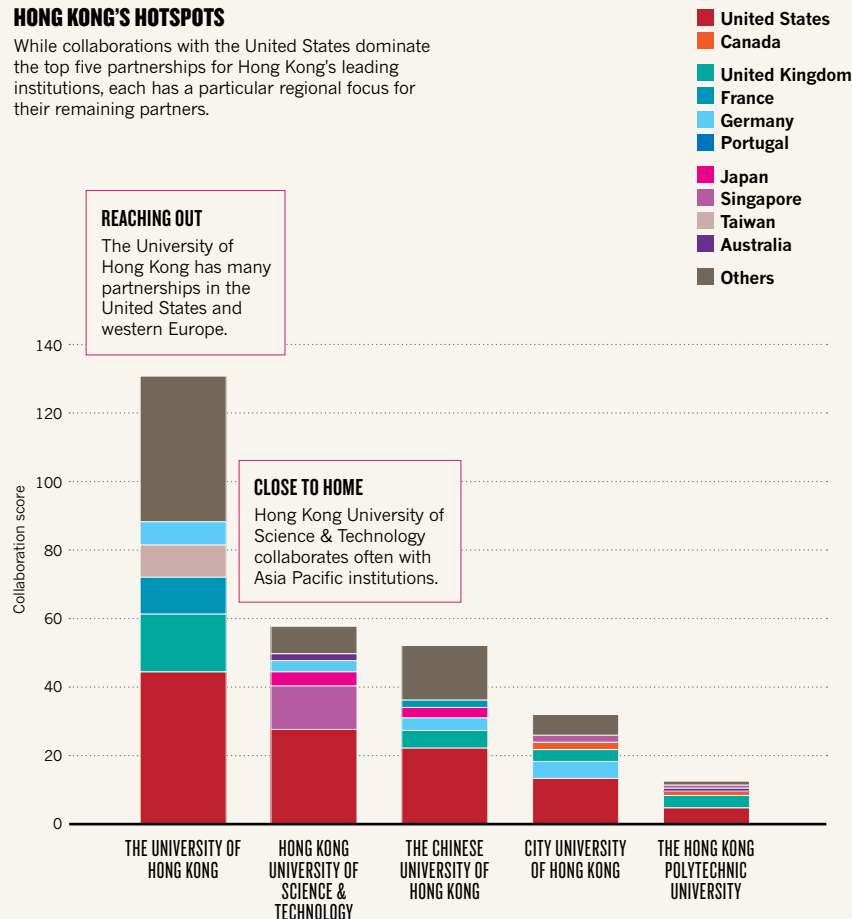
While collaborations with the United States dominate the top five partnerships for Hong Kong's leading institutions, each has a particular regional focus for their remaining partners.

### REACHING OUT

The University of Hong Kong has many partnerships in the United States and western Europe.

### CLOSE TO HOME

Hong Kong University of Science & Technology collaborates often with Asia Pacific institutions.



up of China, a large number of USTC alumni went to overseas institutions, temporarily or permanently. This, combined with a dearth of other local institutes to work with, has pushed the university towards international collaborations for cutting-edge research, particularly in physics and chemistry.

A USTC physics professor, Guo Guoping, says that he and colleagues regularly seek collaborations to promote theoretical developments based on their experimental results. “International partners are crucial to explore our frontier studies,” he says. A good example is a study Guo recently co-authored in *Physics Review Letters* with scientists from the United States and Japan which explores the application of graphene in quantum communication.

**“Support for collaboration is reflected in Hong Kong’s funding policies.”**

USTC is jointly sponsored by CAS and the Ministry of Education. This historic link has contributed to the university's widespread collaborations with CAS institutes. In 2014 CAS was USTC's largest partner, earning it a collaboration score of 136.43. The collaboration score is an indicator of an institution's collaboration in terms of co-authorship of articles in the 68 high-impact journals covered by the index.

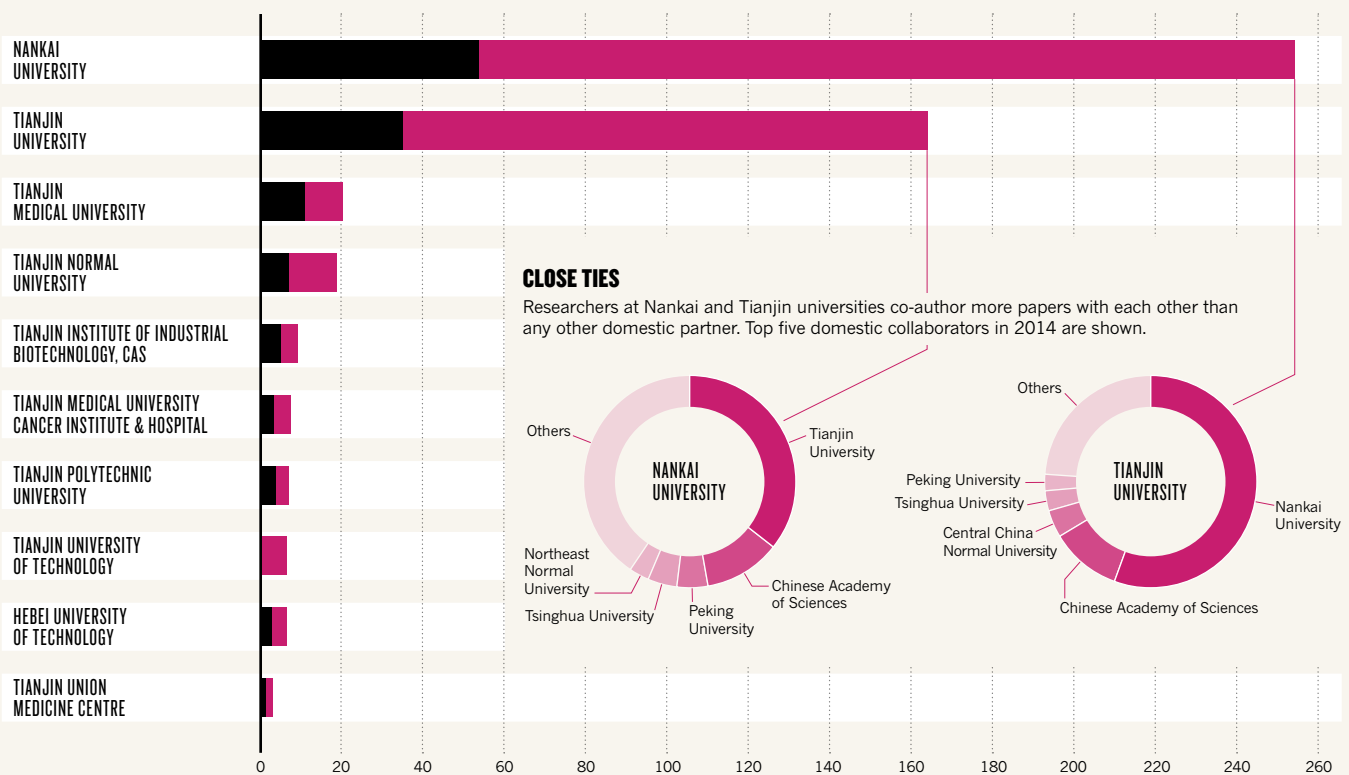
Both Guo and Yuen at HKU argue that while policy support is important, science rather than money drives collaboration.



## IT TAKES TWO

International and domestic collaboration scores of Tianjin's top 10 collaborating institutions.

■ International collaboration score  
■ Domestic collaboration score



"Special collaboration grants to support travel and conferences are good, but only support the original ideas," Guo says.

## TIANJIN

WFC rank China: 10

AC: 461

Tianjin has more universities in the index than Hong Kong and Hefei, although much of its basic research is conducted at two institutions — Nankai and Tianjin universities. These universities had collaboration scores of 254.3 and 163.9 in 2014; much larger than the combined figures of the remaining universities and institutes in this northern city, which neighbours Beijing.

Research at these universities is highly complementary, partly because of historical logistics. In 1952, in line with a Soviet model, the newly-founded People's Republic of China government transferred most of the science departments in Tianjin University (TJU) to Nankai while boosting TJU's engineering capacity. As a result, TJU's engineering research and application is strong, but its basic research is relatively weak. To counter this, the neighbouring universities formed a strong partnership in chemistry and physical science research.

In 2012, partially funded by the Ministry of Education and Tianjin municipal government,

TJU and Nankai jointly established the Tianjin Co-Innovation Center of Chemical Science and Engineering. "The centre has produced many of our co-authorships," says Nankai materials chemist, Yang Zhimou, who along with his team has authored several Nature Index papers.

Many TJU faculty members graduated from Nankai or vice versa and retain strong allegiances to their alma mater. This contributes to ongoing close collaboration, says Ma Jun-An, a TJU Department of Chemistry professor who recently published in *Organic Letters*, a top chemistry journal in the Nature Index.



Tianjin University researchers explore airflow for the first China-developed passenger jet, the C919.

Nankai and TJU are Tianjin's largest collaborators, and in 2014, most frequently collaborated with each other (see 'Close ties'). These strong local links have been forged,

**"International partners are crucial to explore our frontier studies."**

according to Yang, as a result of relatively poor funding and a lack of equipment at Nankai and Tianjin universities. This situation has pushed researchers to exploit resources available locally.

But local partnerships do not thwart collaboration with other domestic and international partners. "Besides strong local partnership, we also have stable collaborations with CAS, CNRS [in France], RIKEN [in Japan] and the University of Texas [in the United States]. They are mutually supportive," Ma says.

The Nankai-TJU Co-innovation Center was established under policies of the ministries of education and of science and technology to encourage collaborations among Chinese research institutions. Initiated in 2011, so far nearly 100 co-innovation centres have been recognized and funded partially by the ministries. Other financial sources come from the universities themselves, local governments and industries, when a centre is industry-oriented. ■

CHINA/GETTY VIA GETTY



# A GUIDE TO THE NATURE INDEX

A description of the terminology and methodology used in this supplement, and a guide to the functionality available free online at [natureindex.com](http://natureindex.com).

The Nature Index is a database of author affiliations and institutional relationships. The index tracks contributions to articles published in a group of highly selective science journals, chosen by an independent group of active researchers.

The Nature Index provides absolute counts of publication productivity at the institutional and national level and, as such, is one indicator of global high-quality research output.

Data in the Nature Index are updated monthly, with the most recent 12 months of data made available under a Creative Commons licence at [natureindex.com](http://natureindex.com).

The database is compiled by Nature Publishing Group (NPG) in collaboration with Digital Science.

The list of journals tracked by the Nature Index is under review, and from 2016 will be extended to include the clinical sciences.

## NATURE INDEX METRICS

There are three measures provided by the Nature Index to track affiliation data. The simplest is the **article count (AC)**. A country or institution is given an AC of 1 for each article that has at least one author from that country or institution. This is the case whether an article has one or a hundred authors, and it means that the same article can contribute to the AC of multiple countries or institutions.

To get a better sense of a country or institution's contribution to an article, and to remove the possibility of counting articles more than once, the Nature Index uses the **fractional count (FC)**, which takes into account the relative contribution of each author to an article. The total FC available per paper is 1, which is shared between all authors under the assumption that each contributed equally. For instance, a paper with 10 authors means that each author receives an FC of 0.1. For authors who have joint affiliations, the individual FC is then split equally between each affiliation.

The third measure used is the **weighted fractional count (WFC)**, which applies a weighting to the FC to adjust for the overrepresentation of papers in astronomy and astrophysics. The four journals in these disciplines publish about 50% of all papers in international journals in this field — approximately five times the equivalent percentage for other fields. Therefore, although the data for astronomy and astrophysics are compiled in the same way as for all other disciplines, articles from these

[natureindex.com](http://natureindex.com) users can search for specific institutions or countries and generate their own reports, ordered by article count (AC), fractional count (FC) or weighted fractional count (WFC).

Each query will return a profile page that lists the country or institution's recent research outputs, from which it is possible to drill down for more information. For example, articles can be displayed by journal, and then by article title. As in the supplement, research outputs are organized by subject area. The profile page also lists the institution or country's top collaborators, as well as its relationship with other research organizations.

journals are assigned one-fifth the weight of other articles (i.e., the FC is multiplied by 0.2 to derive the WFC).

The total FC or WFC for an institution is calculated by summing the FC or WFC for individual authors.

The process is similar for countries, although complicated by the fact that some institutions have overseas labs that will be counted towards their host country totals. What's more, there is great variability in the way authors present their affiliations. Every effort is made to count affiliations consistently, with a background of reasonable assumptions.

For more information on how the affiliation information is processed and counted, please see the FAQ section at [natureindex.com](http://natureindex.com).

## NATUREINDEX.COM

A global indicator of high-quality research

### nature INDEX

Home Institution outputs Country outputs Customer support FAQ

Home / Institution outputs / Institution name

Institution name

Country

Research

Collaboration

Relationships

1 January 2014 - 31 December 2014

Region: Global  
Subject/journal group: All

The table to the right includes counts of all research outputs for Institution name published between 1 January 2014 - 31 December 2014 which are tracked by the Nature Index.

Below, the same research outputs are grouped by subject. Click on the subject to drill-down into a list of articles organized by journal, and then by title.

Note: Articles may be assigned to more than one subject area.

AC	FC	WFC
1221	598.04	558.30

Outputs by subject



Subject	AC	FC	WFC
Chemistry	276	179.1	179.11
Earth & Environmental Sciences	95	42.73	42.73
Life Sciences	439	231.50	231.50
Physical Sciences	652	284.48	244.74

Return to institution outputs

## THE SUPPLEMENT

*Nature Index 2015 China* is based on data from the Nature Index, covering articles published during three consecutive years between 1 January 2012 and 31 December 2014.

Most analyses within the supplement use WFC as the primary metric, as it provides a more even basis for comparison across multiple disciplines, and in determining the relative contribution of each city or institution. Some sections and graphics also refer to collaboration score. This is a relatively new metric that is derived by adding the FC for all the bilateral relationships for that institution or country. If institution A has relationships with two others, B and C, then the collaboration score is the sum of FC for A + B and A + C. ■

# NATURE INDEX CHINA TABLES

China's leading institutions for high-quality science, ordered by weighted fractional count (WFC) for 2014. Also shown are the total number of articles, and the change in WFC from 2013. Articles are from the 68 journals that comprise the Nature Index (see 'How to use the index', S190).

## TOP 200 INSTITUTIONS

2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
1	Peking University (PKU)	275.51	293.86	1,019	6.7%
2	Nanjing University (NJU)	196.52	215.08	518	9.4%
3	Tsinghua University (TH)	195.15	211.39	666	8.3%
4	University of Science and Technology of China (USTC)	175.78	193.90	561	10.3%
5	Zhejiang University (ZJU)	150.44	192.13	364	27.7%
6	Fudan University	129.42	166.21	356	28.4%
7	Institute of Chemistry (ICCAS), CAS	124.85	124.34	306	-0.4%
8	Shanghai Institute of Organic Chemistry (SIOC), CAS	105.62	114.25	210	8.2%
9	Lanzhou University (LZU)	69.72	110.38	186	58.3%
10	Shanghai Jiao Tong University (SJTU)	96.01	108.06	290	12.5%
11	Jilin University (JLU)	97.50	104.93	189	7.6%
12	Wuhan University (WHU)	98.90	96.93	164	-2.0%
13	Xiamen University (XMU)	76.02	95.56	215	25.7%
14	Nankai University (NKU)	113.52	93.43	230	-17.7%
15	Sichuan University (SCU)	76.83	93.36	177	21.5%
16	Soochow University	65.25	91.43	169	40.1%
17	Sun Yat-sen University (SYSU)	79.43	89.72	193	13.0%
18	University of Chinese Academy of Sciences (UCAS)	71.21	89.12	524	25.1%
19	Institute of Physics (IOP), CAS	77.24	87.88	267	13.8%
20	East China Normal University (ECNU)	65.56	83.17	148	26.9%
21	Changchun Institute of Applied Chemistry (CIAC), CAS	80.69	82.09	142	1.7%
22	Hunan University (HNU)	54.57	77.38	111	41.8%
23	Hong Kong University of Science and Technology (HKUST)	54.60	74.62	136	36.7%
24	The University of Hong Kong (HKU)	71.38	71.77	186	0.5%
25	Dalian Institute of Chemical Physics (DICP), CAS	61.90	71.75	139	15.9%
26	East China University of Science and Technology (ECUST)	56.75	71.27	130	25.6%
27	Xi'an Jiaotong University (XJTU)	42.98	67.79	170	57.7%
28	Fujian Institute of Research on the Structure of Matter (FIJISM), CAS	59.54	64.96	124	9.1%
29	Shandong University (SDU)	39.18	63.00	158	60.8%
30	Huazhong University of Science and Technology (HUST)	43.04	57.39	154	33.3%
31	Dalian University of Technology (DUT)	61.42	52.36	96	-14.7%
32	Shanghai Institutes for Biological Sciences (SIBS), CAS	51.44	52.12	131	1.3%
33	Southeast University (SEU)	30.94	51.64	110	66.9%
34	Beijing Normal University (BNU)	39.81	50.82	144	27.7%
35	Northeast Normal University (NENU)	30.73	48.53	67	57.9%
36	Tianjin University (TJU)	33.90	46.23	151	36.4%
37	Tongji University	40.83	45.85	107	12.3%
38	South China University of Technology (SCUT)	30.74	45.79	91	49.0%
39	Hefei Institutes of Physical Science (HIPS), CAS	19.97	39.73	77	99.0%
40	Institute of Semiconductors (IOS), CAS	35.66	36.76	87	3.1%
41	Fuzhou University (FZU)	26.76	35.93	56	34.3%
42	The Chinese University of Hong Kong (CUHK)	39.39	35.82	110	-9.1%
43	People's Liberation Army (PLA)	42.90	35.33	132	-17.6%

2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
44	Harbin Institute of Technology (HIT)	36.22	33.06	63	-8.7%
45	Chinese Academy of Medical Sciences & Peking Union Medical College (CAMS & PUMC)	24.74	32.36	97	30.8%
46	Technical Institute of Physics and Chemistry (TIPC), CAS	29.11	32.08	70	10.2%
47	City University of Hong Kong (CityU)	36.51	31.97	78	-12.4%
48	Shanghai University (SHU)	16.49	31.87	72	93.2%
49	Beijing Institute of Technology (BIT)	20.11	31.35	65	55.9%
50	Beijing University of Chemical Technology (BUCT)	23.45	30.34	50	29.4%
51	Shanghai Institute of Materia Medica (SIMM), CAS	27.48	29.39	56	7.0%
52	University of Science and Technology Beijing (USTB)	25.79	28.85	53	11.9%
53	Shanghai Institute of Ceramics, CAS (SICCAS)	23.82	28.43	65	19.4%
54	Institute of High Energy Physics (IHEP), CAS	32.77	28.02	229	-14.5%
55	Northwest University (NWU)	12.04	26.26	40	118.2%
56	Suzhou Institute of Nano-Tech and Nano-Bionics (SINANO), CAS	25.40	25.94	66	2.1%
57	Institute of Theoretical Physics (ITP), CAS	24.03	24.56	65	2.2%
58	Zhengzhou University (ZZU)	13.77	23.58	53	71.3%
59	Beihang University (BUAA)	17.68	23.36	78	32.1%
60	Southwest University (SWU)	16.11	22.67	37	40.7%
61	Central China Normal University (CCNU)	17.28	22.59	67	30.7%
62	National Center for Nanoscience and Technology (NCNST), CAS	24.20	21.36	42	-11.7%
63	Nanjing Tech University (NanjingTech)	11.92	21.25	49	78.3%
64	Lanzhou Institute of Chemical Physics (LICP), CAS	17.19	20.61	47	19.9%
65	National Astronomical Observatories (NAOC), CAS	21.60	20.08	238	-7.0%
66	Central South University (CSU)	10.18	19.82	58	94.7%
67	Shanghai Institute of Applied Physics (SIAP), CAS	13.02	19.48	58	49.6%
68	The Hong Kong Polytechnic University (PolyU)	25.94	19.40	50	-25.2%
69	Institute of Biophysics (IBP), CAS	18.51	19.36	59	4.6%
70	Chongqing University (CQU)	12.90	19.09	34	48.1%
71	Shandong Normal University (SDNU)	12.73	19.06	24	49.8%
72	University of Electronic Science and Technology of China (UESTC)	11.98	18.42	49	53.7%
73	Institute of Metal Research (IMR), CAS	18.26	17.76	29	-2.7%
74	Shenzhen Institutes of Advanced Technology (SIAT), CAS	11.38	17.51	25	53.8%
75	Kunming Institute of Botany (KIB), CAS	9.98	17.43	29	74.6%
76	China Agricultural University (CAU)	14.65	17.20	44	17.4%
77	China University of Geosciences (CUG)	14.52	17.01	38	17.2%
78	Nanjing Normal University (NNU)	7.75	16.71	45	115.6%
79	Henan Normal University (HTU)	12.42	16.51	38	32.9%
80	State Oceanic Administration (SOA)	6.71	16.14	47	140.7%
81	Wuhan University of Technology (WUT)	8.59	16.05	29	86.9%
82	Shanghai Institute of Technical Physics (SITP), CAS	7.97	15.94	54	100.0%
83	Yunnan University (YNU)	11.39	15.80	36	38.6%
84	South China Normal University (SCNU)	6.87	15.72	35	128.7%
85	Institute of Zoology (IOZ), CAS	15.43	15.65	40	1.4%
86	Ocean University of China (OUC)	13.03	15.59	40	19.6%
87	Xiangtan University (XTU)	12.87	15.50	29	20.4%
88	Wuhan Institute of Physics and Mathematics (WIPM), CAS	10.35	15.34	53	48.3%
89	Shanghai Institute of Microsystem and Information Technology (SIMIT), CAS	17.08	15.21	30	-10.9%
90	Nanjing Medical University (NJMU)	7.27	14.91	58	105.1%
91	Shanxi University (SXU)	10.82	14.66	29	35.6%
92	BGI	15.71	14.66	63	-6.7%
93	Huazhong Agricultural University (HZAU)	10.67	14.38	31	34.8%
94	Changzhou University (CZU)	5.93	14.23	22	140.2%
95	China University of Petroleum (UPC)	7.15	14.17	37	98.2%
96	Northwestern Polytechnical University (NPU)	12.90	13.79	28	6.9%



2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
97	Northeastern University (NEU)	2.10	13.66	21	550.0%
98	Institute of Genetics and Developmental Biology (IGDB), CAS	10.20	13.43	38	31.7%
99	Ningbo Institute of Materials Technology and Engineering (NIMTE), CAS	14.94	13.28	23	-11.1%
100	Beijing University of Technology (BJUT)	7.67	13.27	29	73.0%
101	Shanghai Normal University (SHNU)	3.75	13.20	24	251.8%
102	China Academy of Engineering Physics (CAEP)	13.27	13.11	50	-1.2%
103	Qingdao University of Science and Technology (QUST)	10.66	13.08	25	22.7%
104	Shaanxi Normal University (SNNU)	11.80	12.95	22	9.8%
105	Nanjing University of Aeronautics and Astronautics (NUAA)	6.58	12.84	26	95.0%
106	Anhui Normal University (AHNU)	5.96	12.80	15	114.8%
107	Hong Kong Baptist University (HKBU)	12.77	12.78	33	0.1%
108	South University of Science and Technology of China (SUSTC)	2.24	12.52	31	458.3%
109	China Meteorological Administration (CMA)	9.93	12.34	35	24.3%
110	Wenzhou University (WZU)	8.05	11.97	21	48.6%
111	Research Center for Eco-Environmental Sciences (RCEES), CAS	6.60	11.92	20	80.5%
112	South China Sea Institute of Oceanology (SCSIO), CAS	10.59	11.82	23	11.6%
113	Institute of Oceanology, CAS (IOCAS)	5.12	11.48	22	124.4%
114	Henan University (HENU)	12.53	11.35	18	-9.4%
115	Institute of Atmospheric Physics (IAP), CAS	17.87	11.34	38	-36.5%
116	Zhejiang University of Technology (ZJUT)	8.03	11.10	22	38.2%
117	China Pharmaceutical University (CPU)	3.88	10.87	24	180.1%
118	National University of Defense Technology (NUDT)	16.16	10.68	44	-33.9%
119	National Institute of Biological Sciences, Beijing (NIBS)	11.73	10.65	31	-9.2%
120	Jinan University (JNU)	4.31	10.62	30	146.3%
121	Hunan Normal University (HUNNU)	11.12	10.38	17	-6.7%
122	Nanchang University (NCU)	10.06	9.75	22	-3.1%
123	Xidian University	3.78	9.65	12	155.5%
124	Institute of Microbiology (IM), CAS	9.09	9.59	24	5.6%
125	Nanjing University of Posts and Telecommunications (NUPT)	8.28	9.54	21	15.2%
126	Institute of Geology and Geophysics (IGG), CAS	7.94	9.48	22	19.3%
127	Renmin University of China (RUC)	8.89	9.46	24	6.4%
128	China Earthquake Administration (CEA)	9.50	9.46	26	-0.5%
129	Shanghai Institute of Optics and Fine Mechanics (SIOM), CAS	9.82	9.19	17	-6.4%
130	Qingdao Institute of Bioenergy and Bioprocess Technology (QIBET), CAS	2.90	9.10	19	213.9%
131	Guangzhou Institutes of Biomedicine and Health (GIBH), CAS	19.60	8.91	18	-54.6%
132	Jiangsu Normal University (JSNU)	5.61	8.66	13	54.4%
133	Hebei University (HBU)	7.86	8.11	13	3.1%
134	Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), CAS	10.33	7.95	16	-23.1%
135	University of Shanghai for Science and Technology (USST)	4.15	7.94	12	91.5%
136	Hangzhou Normal University (HZNU)	9.52	7.94	33	-16.6%
137	Shenzhen University (SZU)	5.05	7.73	26	53.0%
138	Jiangxi Normal University (JXNU)	3.14	7.71	23	146.0%
139	Hefei University of Technology (HFUT)	8.97	7.70	16	-14.1%
140	Nanjing University of Science and Technology (NUST)	6.43	7.60	18	18.3%
141	Yangzhou University (YZU)	4.67	7.47	13	60.1%
142	Anhui University (AHU)	3.71	7.28	18	96.1%
143	Zhejiang Normal University (ZJNU)	8.44	7.27	15	-13.9%
144	Northwest A&F University (NWAUFU)	9.90	7.26	19	-26.6%
145	Shantou University (STU)	7.13	6.99	18	-1.9%
146	Institute of Botany (IBCAS)	3.02	6.71	15	122.3%
147	Southern Medical University (SMU)	3.47	6.70	23	93.4%
148	Chinese Academy of Agricultural Sciences (CAAS)	8.10	6.66	26	-17.8%
149	Hebei Normal University	1.40	6.59	15	371.3%

2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
150	Purple Mountain Observatory (PMO), CAS	6.31	6.40	97	1.5%
151	Heilongjiang University (HLJU)	6.67	6.39	10	-4.2%
152	North China Electric Power University (NCEPU)	6.50	6.27	14	-3.6%
153	Jiangnan University (JU)	8.91	6.21	12	-30.3%
154	Nanjing University of Information Science and Technology (NUIST)	7.74	6.09	25	-21.3%
155	Jiangsu University (JU)	3.42	6.04	17	76.4%
156	Beijing Jiaotong University (BJTU)	4.15	5.97	15	43.9%
157	Nanjing Agricultural University (NAU)	2.59	5.83	18	125.0%
158	Northwest Normal University (NWNNU)	0.25	5.71	11	2,184.8%
159	Xinjiang Technical Institute of Physics and Chemistry (XTIPC), CAS	4.98	5.66	10	13.8%
160	University of Jinan (UJN)	5.57	5.65	10	1.4%
161	Capital Medical University (CMU)	3.05	5.54	38	81.6%
162	Shanghai Astronomical Observatory (SHAO), CAS	4.76	5.51	102	15.8%
163	Xuzhou Medical University (XZMC)	0.31	5.39	14	1,665.9%
164	Guangxi Normal University (GXNU)	4.60	5.30	13	15.3%
165	Tianjin Medical University (TMC)	6.73	5.21	17	-22.7%
166	Yantai Institute of Coastal Zone Research (YIC), CAS	2.14	4.95	8	131.2%
167	Capital Normal University (CNU)	6.65	4.88	19	-26.6%
168	Institute of Tibetan Plateau Research (ITP), CAS	7.22	4.73	16	-34.5%
169	Tianjin Normal University (TJNU)	0.94	4.71	18	402.1%
170	Qufu Normal University (QFNU)	3.63	4.67	9	28.8%
171	Harbin Engineering University (HEU)	2.52	4.67	7	85.1%
172	Zhejiang Sci-Tech University (ZSTU)	4.94	4.64	13	-6.2%
173	Institute of Microelectronics (IME), CAS	6.80	4.44	11	-34.7%
174	Chengdu Institute of Biology (CIB), CAS	5.45	4.44	14	-18.6%
175	Beijing University of Posts and Telecommunications (BUPT)	3.42	4.37	7	28.0%
176	BGI	3.82	4.35	21	14.0%
177	Linyi University (LYU)	3.19	4.29	18	34.4%
178	Guizhou University (GZU)	3.18	4.29	10	35.0%
179	Ningbo University (NBU)	7.56	4.27	14	-43.5%
180	Qingdao University (QU)	2.89	4.27	14	47.6%
181	Institute of Vertebrate Paleontology and Paleoanthropology (IVPP), CAS	4.19	4.27	20	1.8%
182	Hebei University of Technology (HEBUT)	1.65	4.25	7	158.0%
183	Institute of Process Engineering (IPE), CAS	3.57	4.22	14	18.1%
184	Huaibei Normal University (HUN)	4.78	4.19	9	-12.2%
185	Yanshan University (YSU)	4.03	4.14	7	2.8%
186	Yantai University	0.64	4.12	9	541.1%
187	Institute of Mechanics (IM), CAS	3.84	4.09	9	6.7%
188	Tianjin University of Technology (TUT)	5.53	4.00	7	-27.7%
189	Chongqing Medical University (CQMU)	4.31	3.83	14	-11.2%
190	Wenzhou Medical University (WMU)	1.49	3.81	11	155.7%
191	Kunming University of Science and Technology (KUST)	3.04	3.77	9	24.1%
192	Chinese Center for Disease Control and Prevention (China CDC)	3.20	3.77	24	17.7%
193	Chengdu Institute of Organic Chemistry (CIOC), CAS	4.16	3.76	8	-9.6%
194	Donghua University (DHU)	6.34	3.72	11	-41.4%
195	Henan Polytechnic University (HPU)	0.68	3.71	13	442.5%
196	Anhui Medical University (AHMU)	2.12	3.71	16	74.7%
197	Hubei University (HUBU)	3.63	3.68	7	1.4%
198	National Space Science Center (NSSC), CAS	2.87	3.65	28	27.3%
199	Institute of Coal Chemistry (ICC), CAS	1.69	3.61	9	113.8%
200	Institute of Electrical Engineering (IEE), CAS	2.09	3.60	7	72.3%

## TOP 50 INSTITUTIONS IN LIFE SCIENCES

2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
1	Shanghai Institutes for Biological Sciences (SIBS), CAS	49.30	51.12	125	3.7%
2	Peking University (PKU)	41.54	47.77	156	15.0%
3	Tsinghua University (TH)	27.71	28.23	110	1.9%
4	Shanghai Jiao Tong University (SJTU)	20.50	27.31	92	33.2%
5	Zhejiang University (ZJU)	16.53	23.03	75	39.4%
6	People's Liberation Army (PLA)	26.79	22.02	94	-17.8%
7	Sun Yat-sen University (SYSU)	11.21	21.47	60	91.5%
8	Chinese Academy of Medical Sciences & Peking Union Medical College (CAMS & PUMC)	13.44	19.02	70	41.5%
9	Shandong University (SDU)	6.63	18.28	40	175.7%
10	Fudan University	21.72	17.72	89	-18.4%
11	University of Science and Technology of China (USTC)	19.07	17.52	45	-8.1%
12	Institute of Biophysics (IBP), CAS	14.39	17.36	54	20.6%
13	University of Chinese Academy of Sciences (UCAS)	8.36	14.74	92	76.2%
14	Institute of Zoology (IOZ), CAS	15.43	14.19	37	-8.0%
15	BGI	15.56	13.75	62	-11.7%
16	Institute of Genetics and Developmental Biology (IGDB), CAS	10.18	13.31	37	30.8%
17	Wuhan University (WHU)	15.08	12.91	29	-14.4%
18	The University of Hong Kong (HKU)	10.28	12.49	46	21.5%
19	Nanjing University (NJU)	9.32	11.40	35	22.3%
20	Huazhong University of Science and Technology (HUST)	8.71	10.59	42	21.6%
21	Huazhong Agricultural University (HZAU)	7.61	10.03	23	31.9%
22	Hong Kong University of Science and Technology (HKUST)	5.48	9.85	18	79.7%
23	National Institute of Biological Sciences, Beijing (NIBS)	9.77	9.38	26	-4.0%
24	Xiamen University (XMU)	5.83	9.16	31	57.1%
25	China Agricultural University (CAU)	9.39	9.12	29	-2.9%
26	Beijing Normal University (BNU)	5.38	8.49	30	57.9%
27	Institute of Microbiology (IM), CAS	5.65	8.37	22	48.2%
28	Nanjing Medical University (NJMU)	5.86	8.18	30	39.7%
29	Nankai University (NKU)	8.46	8.09	18	-4.3%
30	East China Normal University (ECNU)	6.73	7.81	19	16.0%
31	The Chinese University of Hong Kong (CUHK)	6.44	7.70	33	19.5%
32	Sichuan University (SCU)	2.15	7.13	38	231.4%
33	Tongji University	8.01	6.70	36	-16.4%
34	Soochow University	5.12	6.35	22	24.1%
35	Institute of Botany (IBCAS)	2.97	5.68	13	90.8%
36	Tianjin Medical University (TMC)	5.79	5.21	17	-10.1%
37	Shanghai Institute of Materia Medica (SIMM), CAS	10.26	4.99	18	-51.3%
38	Central South University (CSU)	1.72	4.11	21	138.5%
39	Beijing Institute of Genomics (BIG), CAS	3.61	4.03	19	11.9%
40	Capital Medical University (CMU)	2.93	4.02	32	37.5%
41	Institute of Vertebrate Paleontology and Paleoanthropology (IVPP), CAS	3.96	3.95	18	-0.2%
42	Southeast University (SEU)	2.51	3.91	11	55.9%
43	Chinese Academy of Agricultural Sciences (CAAS)	6.53	3.63	17	-44.4%
44	Wenzhou Medical University (WMU)	1.08	3.49	10	224.7%
45	Southern Medical University (SMU)	2.77	3.48	14	26.0%
46	Anhui Medical University (AHMU)	2.12	3.41	14	61.0%
47	Yunnan University (YNU)	1.11	3.30	10	196.1%
48	Xi'an Jiaotong University (XJTU)	0.81	3.27	16	303.9%
49	Kunming Institute of Zoology (KIZ), CAS	3.19	3.21	16	0.6%
50	Nanjing Agricultural University (NAU)	2.47	3.13	12	26.8%



## TOP 50 INSTITUTIONS IN CHEMISTRY

2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
1	Peking University (PKU)	142.60	152.25	378	6.8%
2	Nanjing University (NJU)	117.69	129.85	229	10.3%
3	Zhejiang University (ZJU)	86.44	129.11	189	49.4%
4	Institute of Chemistry (ICCAS), CAS	119.81	120.18	287	0.3%
5	Fudan University	80.30	117.15	198	45.9%
6	Shanghai Institute of Organic Chemistry (SIOC), CAS	105.23	113.47	206	7.8%
7	University of Science and Technology of China (USTC)	93.78	111.24	233	18.6%
8	Tsinghua University (TH)	92.19	107.43	249	16.5%
9	Lanzhou University (LZU)	50.35	88.25	124	75.3%
10	Jilin University (JLU)	73.64	80.64	133	9.5%
11	Changchun Institute of Applied Chemistry (CIAC), CAS	79.52	79.25	134	-0.3%
12	Sichuan University (SCU)	68.45	75.15	107	9.8%
13	Xiamen University (XMU)	59.87	73.56	146	22.9%
14	Dalian Institute of Chemical Physics (DICP), CAS	61.65	70.08	135	13.7%
15	Hunan University (HNU)	51.90	69.09	91	33.1%
16	Nankai University (NKU)	86.00	67.62	166	-21.4%
17	East China University of Science and Technology (ECUST)	53.04	66.33	112	25.0%
18	Wuhan University (WHU)	63.32	65.98	100	4.2%
19	Fujian Institute of Research on the Structure of Matter (FJIRSM), CAS	55.22	63.42	121	14.9%
20	Soochow University	34.69	62.89	104	81.3%
21	Sun Yat-sen University (SYSU)	52.10	56.43	97	8.3%
22	University of Chinese Academy of Sciences (UCAS)	47.89	55.93	278	16.8%
23	East China Normal University (ECNU)	41.54	55.26	86	33.0%
24	Shanghai Jiao Tong University (SJTU)	45.14	51.63	93	14.4%
25	Northeast Normal University (NENU)	25.65	42.59	54	66.0%
26	Dalian University of Technology (DUT)	47.16	41.33	76	-12.4%
27	South China University of Technology (SCUT)	28.75	39.39	73	37.0%
28	Tianjin University (TJU)	25.59	37.21	131	45.4%
29	The University of Hong Kong (HKU)	39.04	36.31	60	-7.0%
30	Hong Kong University of Science and Technology (HKUST)	25.86	36.24	72	40.2%
31	Shandong University (SDU)	19.69	31.88	57	61.9%
32	Fuzhou University (FZU)	26.51	31.23	49	17.8%
33	Southeast University (SEU)	9.78	28.73	57	193.7%
34	Technical Institute of Physics and Chemistry (TIPC), CAS	22.50	28.42	61	26.3%
35	Beijing University of Chemical Technology (BUCT)	20.35	28.33	46	39.2%
36	Tongji University	19.51	28.04	44	43.7%
37	Huazhong University of Science and Technology (HUST)	12.58	26.68	58	112.0%
38	Xi'an Jiaotong University (XJTU)	10.72	26.50	67	147.1%
39	Shanghai Institute of Materia Medica (SIMM), CAS	18.28	24.56	39	34.4%
40	Northwest University (NWU)	10.29	21.95	25	113.3%
41	Shanghai University (SHU)	8.57	21.54	43	151.1%
42	Beijing Normal University (BNU)	10.56	21.11	37	99.9%
43	Hefei Institutes of Physical Science (HIPS), CAS	8.68	20.92	40	141.1%
44	Southwest University (SWU)	12.51	20.67	26	65.2%
45	Beijing Institute of Technology (BIT)	16.95	20.47	42	20.7%
46	Institute of Physics (IOP), CAS	14.31	20.41	75	42.7%
47	Nanjing Tech University (NanjingTech)	11.04	19.81	42	79.4%
48	National Center for Nanoscience and Technology (NCNST), CAS	21.82	19.69	38	-9.8%
49	Harbin Institute of Technology (HIT)	14.15	19.20	35	35.7%
50	Lanzhou Institute of Chemical Physics (LICP), CAS	15.25	18.86	44	23.7%

## TOP 50 INSTITUTIONS IN PHYSICAL SCIENCES

2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
1	Peking University (PKU)	105.05	111.30	525	6.0%
2	Tsinghua University (TH)	87.99	97.52	370	10.8%
3	Institute of Physics (IOP), CAS	72.08	77.59	233	7.6%
4	University of Science and Technology of China (USTC)	71.44	73.68	310	3.1%
5	Nanjing University (NJU)	65.06	69.53	247	6.9%
6	Fudan University	39.34	56.56	116	43.8%
7	Zhejiang University (ZJU)	64.83	48.59	126	-25.1%
8	Xi'an Jiaotong University (XJTU)	32.80	39.81	93	21.4%
9	Institute of Semiconductors (IOS), CAS	35.66	36.01	82	1.0%
10	Shanghai Jiao Tong University (SJTU)	34.28	34.57	118	0.8%
11	Hong Kong University of Science and Technology (HKUST)	22.12	33.39	61	50.9%
12	Soochow University	27.26	30.52	63	11.9%
13	Jilin University (JLU)	28.87	28.63	59	-0.8%
14	Institute of Chemistry (ICCAS), CAS	25.51	27.88	68	9.3%
15	Huazhong University of Science and Technology (HUST)	24.74	25.24	69	2.0%
16	Southeast University (SEU)	20.09	25.01	57	24.5%
17	Institute of Theoretical Physics (ITP), CAS	24.03	23.78	63	-1.0%
18	The University of Hong Kong (HKU)	25.81	21.90	77	-15.2%
19	Nankai University (NKU)	23.56	21.69	55	-7.9%
20	East China Normal University (ECNU)	16.97	21.32	47	25.6%
21	National Astronomical Observatories (NAOC), CAS	20.84	19.80	235	-5.0%
22	University of Science and Technology Beijing (USTB)	21.29	18.75	34	-12.0%
23	Wuhan University (WHU)	18.52	18.71	40	1.0%
24	Shanghai Institute of Ceramics, CAS (SICCAS)	19.59	18.57	42	-5.2%
25	Hefei Institutes of Physical Science (HIPS), CAS	11.26	18.19	35	61.5%
26	Sun Yat-sen University (SYSU)	21.90	17.94	44	-18.1%
27	University of Chinese Academy of Sciences (UCAS)	16.43	17.80	161	8.3%
28	Lanzhou University (LZU)	13.15	17.16	48	30.4%
29	Institute of High Energy Physics (IHEP), CAS	24.11	17.12	209	-29.0%
30	University of Electronic Science and Technology of China (UESTC)	11.41	17.06	41	49.5%
31	Beijing Normal University (BNU)	19.18	16.96	58	-11.6%
32	Beijing Institute of Technology (BIT)	7.84	16.89	33	115.4%
33	Harbin Institute of Technology (HIT)	22.25	16.07	36	-27.8%
34	Shanghai Institute of Technical Physics (SITP), CAS	7.97	15.13	48	89.8%
35	Suzhou Institute of Nano-Tech and Nano-Bionics (SINANO), CAS	13.85	14.57	38	5.2%
36	City University of Hong Kong (CityU)	20.31	14.51	42	-28.5%
37	Beihang University (BUAA)	10.07	14.13	59	40.3%
38	Shandong University (SDU)	14.14	13.81	64	-2.3%
39	Sichuan University (SCU)	7.74	13.02	40	68.2%
40	Tongji University	13.18	12.52	29	-5.0%
41	Nanjing University of Aeronautics and Astronautics (NUAA)	4.58	12.32	19	168.7%
42	Hunan University (HNU)	4.32	12.18	25	181.9%
43	Shanghai Institute of Microsystem and Information Technology (SIMIT), CAS	15.85	12.17	26	-23.2%
44	Northwestern Polytechnical University (NPU)	11.74	12.11	22	3.2%
45	Dalian University of Technology (DUT)	17.98	11.56	22	-35.7%
46	Changchun Institute of Applied Chemistry (CIAC), CAS	11.07	11.55	20	4.3%
47	The Chinese University of Hong Kong (CUHK)	13.12	11.45	37	-12.7%
48	Xiamen University (XMU)	11.91	11.33	38	-4.9%
49	National University of Defense Technology (NUDT)	16.07	10.29	40	-36.0%
50	Shanghai University (SHU)	8.28	10.23	28	23.6%

## TOP 25 INSTITUTIONS IN EARTH AND ENVIRONMENTAL SCIENCES

2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
1	Nanjing University (NJU)	9.72	14.94	33	53.7%
2	Peking University (PKU)	7.31	14.71	50	101.1%
3	State Oceanic Administration (SOA)	5.90	14.67	32	148.5%
4	China Meteorological Administration (CMA)	9.76	12.32	34	26.2%
5	Institute of Atmospheric Physics (IAP), CAS	17.79	11.06	36	-37.8%
6	China University of Geosciences (CUG)	10.40	10.96	26	5.3%
7	Institute of Oceanology, CAS (IOCAS)	5.12	10.03	18	96.0%
8	South China Sea Institute of Oceanology (SCSIO), CAS	5.41	9.88	20	82.7%
9	China Earthquake Administration (CEA)	9.50	9.46	26	-0.5%
10	Ocean University of China (OUC)	9.60	8.92	24	-7.1%
11	University of Science and Technology of China (USTC)	6.91	8.71	22	26.2%
12	Institute of Geology and Geophysics (IGG), CAS	7.54	8.26	19	9.5%
13	Beijing Normal University (BNU)	6.78	7.58	27	11.8%
14	University of Chinese Academy of Sciences (UCAS)	3.22	5.44	30	69.0%
15	Xiamen University (XMU)	1.81	5.33	14	194.1%
16	Lanzhou University (LZU)	5.25	5.29	15	0.9%
17	Nanjing University of Information Science and Technology (NUIST)	6.37	4.96	22	-22.1%
18	Institute of Tibetan Plateau Research (ITP), CAS	7.22	4.73	16	-34.5%
19	Wuhan University (WHU)	5.93	4.65	9	-21.5%
20	Hong Kong University of Science and Technology (HKUST)	4.54	4.46	5	-1.7%
21	Tsinghua University (TH)	3.09	3.89	13	25.9%
22	Institute of Earth Environment (IEE), CAS	2.66	3.50	11	32.0%
23	Guangzhou Institute of Geochemistry (GIG), CAS	3.82	3.49	11	-8.4%
24	Zhejiang University (ZJU)	2.05	3.26	8	59.2%
25	Yunnan University (YNU)	1.12	3.14	6	181.2%

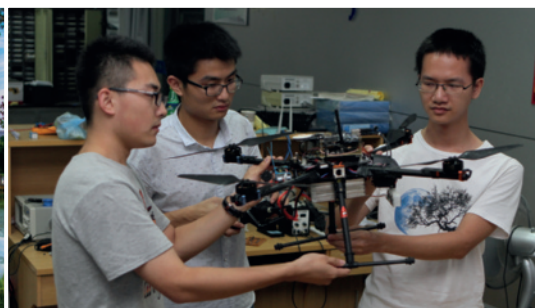
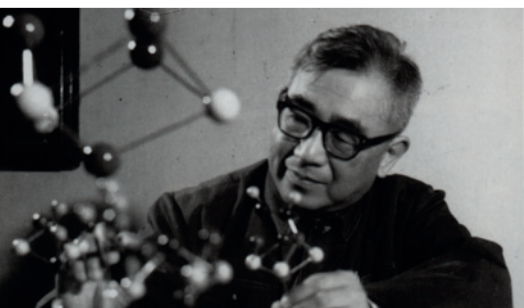
## TOP 25 INSTITUTIONS IN NATURE AND SCIENCE

2014	INSTITUTION	WFC 2013	WFC 2014	AC 2014	CHANGE IN WFC 2013-2014
1	Peking University (PKU)	4.10	6.48	28	58.0%
2	Tsinghua University (TH)	5.43	4.90	20	-9.8%
3	BGI	1.78	2.84	14	59.1%
4	Shanghai Institutes for Biological Sciences (SIBS), CAS	5.44	2.70	10	-50.4%
5	Institute of Biophysics (IBP), CAS	0.13	2.56	7	1,856.6%
6	Zhejiang University (ZJU)	1.70	2.19	8	28.9%
7	National Institute of Biological Sciences, Beijing (NIBS)	1.12	2.11	4	87.3%
8	Dalian Institute of Chemical Physics (DICP), CAS	0.89	1.56	4	75.7%
9	University of Chinese Academy of Sciences (UCAS)	0.50	1.20	7	138.5%
10	Ocean University of China (OUC)	0.17	1.17	3	600.0%
11	Institute of Physics (IOP), CAS	0.95	1.12	5	18.0%
12	Huazhong University of Science & Technology (HUST)	0.05	0.92	1	1,733.3%
13	Shanghai Institute of Organic Chemistry (SIOC), CAS		0.92	1	
14	Yanshan University	0.73	0.82	1	11.6%
15	University of Science and Technology of China (USTC)	1.69	0.82	7	-51.6%
16	Harbin Institute of Technology (HIT)	0.07	0.80	1	1,020.0%
17	China National Genebank (CNCB)		0.64	7	
18	Second Military Medical University (SMMU)		0.64	3	
19	Nanjing University (NJU)	0.05	0.58	2	1,008.3%
20	Chinese Academy of Agricultural Sciences (CAAS)	2.84	0.56	5	-80.5%
21	Shanghai Jiao Tong University (SJTU)	0.18	0.54	1	193.1%
22	Nanjing Institute of Geology and Palaeontology, CAS	0.75	0.53	2	-29.5%
23	Chinese Academy of Medical Sciences & Peking Union Medical College (CAMS & PUMC)	0.13	0.53	3	322.5%
24	Yunnan University	0.80	0.50	1	-37.5%
25	Institute of Oceanology, CAS (IOCAS)		0.50	1	

Weighted fractional count (WFC) for each institution is shown to two decimal places only. When two or more institutions have the same WFC, their positions are determined by the thousandth place (or beyond).

These results are based on the most recent data available as of 14 September 2015. Owing to continual refinements of the data, the figures in the database are liable to change and might differ to those printed in the supplements.





Fuzhou University

# Engine promoting innovation in Southeast China

**Founded in 1958, Fuzhou University is in Fuzhou city, the capital of Fujian Province in southeast China. It is one of a 100 universities selected as part of the Ministry of Education's prestigious 211 Project aimed at strengthening higher education and scientific research.**

**F**uzhou University has been listed in the global top 1 per cent for chemistry, engineering and materials science research by Thomson Reuters' Essential Science Indicators (ESI) 2014. The university was ranked 36 in Nature Index 2014 China and 23 in ESI's top 100 highly cited Chinese universities. In 2014, eight Fuzhou professors featured in Elsevier's Most Cited Chinese Researchers. Some of the university's latest research is highlighted below.

## Chemistry

**Photocatalysis.** Fuzhou University's Research Institute of Photocatalysis, initiated by Xianzhi Fu in 1997, became a State Key Laboratory in 2013. It focuses on searching for new types of photocatalysts and co-catalysts for a range of applications. It has won one second-class Chinese National Award for Advancement in Science and Technology, two first-class, provincial-level awards within the same category, and one first-class award from the People's Liberation Army. In 2009, Fu was elected as an academician of the Chinese Academy of Engineering (CAE).

## Industrial catalysis and biological analysis.

The National Engineering Research Center of Chemical Fertilizer Catalysts (NERC-CFC) was founded by Kemei Wei, a CAE academician.

Focused on environmentally friendly catalysts for ammonia and hydrogen production, exhaust gas treatment and clean fuel production, NERC-CFC has won five national awards, and seven provincial and ministerial awards.

The Key Laboratory of Analysis and Detection Technology for Food Safety of the Ministry of Education (MOE) explores electrochemiluminescence bioanalysis, nanobiosensors and biomarker analysis in living biological systems. The laboratory has been awarded nine provincial and ministerial awards.

## Materials science

Jiaxi Lu, a key founding member of Fuzhou University, established the field of crystalline materials science. Over the past six decades, the field has become an influential research subject in China and one of the distinctive disciplines of Fuzhou University. It focuses on the synthesis of diverse crystalline materials, the relationship between structures and properties, and the application of specific crystalline materials to magneto-optics, lasers, and nonlinear optics.

Researchers at the Institute of Advanced Energy Materials have developed a new synthetic strategy for the self-assembly of mesostructural materials with controllable crystal phases, facets, dimensions, sizes, pores and morphologies, and discovered the relationships between the intrinsic characteristics of mesostructural materials and their photovoltaic and electrochemical properties.

The biomedical materials research group focuses on novel biomedical materials and their applications in diagnostics, theranostics, tissue engineering and biosimulation.

## Physics

The Laboratory of Quantum Optics, led by Shibiao Zheng, a Yangtze River Scholar Professor, has proposed many important cavity-quantum-electrodynamics-based schemes for realizing entanglement and quantum logic operations. Of their papers published in Physics Review Letters, one has been cited over 700 times in Thomson Reuters's Science Citation Index journals. Zheng won the second-class National Award for Natural Science, and the National Award for Youth in Science and Technology.

The National Engineering Laboratory for Flat Panel Display focuses on the design, preparation and performance optimization of novel photoelectronic devices, such as printing displays and light-harvesting devices.

## Mathematics and computer science

The Center for Discrete Mathematics and Theoretical Computer Science (DIMACS-FU), headed by Genghua Fan, became the MOE Key Laboratory of Discrete Mathematics with Applications in 2007. Its research focuses on graph theory and combinatorics, mathematical methods in very large-scale integration.



## Contact

**Visit:** <http://www.fzu.edu.cn/>

**Fax:** +86-0591-22866099

**E-mail:** [faomail@fzu.edu.cn](mailto:faomail@fzu.edu.cn)

**Address:** No.2 Xue yuan Road, Minhou, Fuzhou, Fujian, China, 350116



# :insideview

## profile feature



**Zhou Guanghong**, *president of Nanjing Agricultural University.*

With a history that extends back to 1902, Nanjing Agricultural University (NAU) is one of China's oldest higher education institutions with an agricultural science focus. Now one of the top 50 universities in the country, NAU has matured into a comprehensive research university with 20 colleges offering hundreds of degree programmes and promoting cutting-edge research in basic and applied sciences. Here, Zhou Guanghong, president of NAU, discusses the development of the university and his vision for building a world-renowned institution for agricultural education and research.

### **Q. What are NAU's primary development goals for the next ten years?**

Our goal is to become one of the world's top universities for agricultural sciences. To accomplish this, we must attract top-level researchers, provide world-class training to our students, advance our research in the agricultural, life and environmental sciences, and — crucially in my view — conduct innovative research that benefits society. NAU already ranks 78th in the 2015 National Taiwan University Ranking for the field of agriculture. By 2020, we hope to be among the top 50 universities globally in our core subject areas, and we want to be among the top 500 universities in the Academic Ranking of World Universities in all subject areas by 2030.

We recognize the importance of aligning our development goals with national interests. Specifically, our research should not be limited to agriculture and crop production, but it must also focus on the impact of agricultural and rural development as well as on the coordinated development of food security along with animal and human health.

### **Q. How do you plan to achieve these goals?**

We have initiated several programmes to enhance the growth of our researchers, including the Zhongshan Scholar Project, which supports career development and encourages innovation from outstanding scientists selected for the programme. In addition, we have established a postdoctoral faculty track that enables us to identify the most promising

scholars for faculty positions. To promote professional competence, we have adopted international standards to recruit and evaluate faculty members.

In addition to attracting top scholars to boost the strength of our research, we are reinforcing our core subject areas. We are establishing novel research platforms and building two new campuses to ensure we will have adequate space and research facilities as we grow.

We recognize that international cooperation must play a key role as we evolve into a world-class university. NAU maintains ties with over 160 universities and research institutes around the world, and has established 10 international research centres. We participate in many international collaborative research programmes and are continuously augmenting academic exchanges with world-leading institutions. We are now working on setting up more international exchange programmes for graduate students.

“

The soul of a university lies in its academic spirit, but a university also has the broader mission to serve all of society.

”

### **Q. What are NAU's strengths?**

NAU is located in Nanjing, one of China's great ancient capitals and the modern and vibrant capital city of Jiangsu Province. The university shares a proud historical tradition with the city as it was the first Chinese university to offer four-year bachelor degrees in the agricultural sciences. Today, as a national key university under the Chinese Ministry of Education, we balance that heritage with modern expertise and leadership. In the area of agricultural sciences, NAU currently ranks in the top 0.1 percent globally according to Thomson Reuters Essential Science Indicators, while we are ranked in the top 1 per cent in the areas of plant and animal science, environment and ecology,

and biology and biochemistry. In our core competencies, notable results from research at NAU in crop genetics and breeding, crop growth modelling, quality control and processing of agricultural products, pest control, veterinary medicine, utilization of agricultural waste, and bioorganic fertilizers have been published in prominent international journals.

### **Q. How can academic research and student education contribute to each other?**

Academic research and student training certainly go hand-in-hand and should develop concurrently. NAU encourages its undergraduate students to gain hands-on research skills by working directly on research projects. Our well-established research platforms and experienced faculty at NAU also give students the opportunity to see leading-edge research firsthand, giving them a better understanding of how their research fields are advancing. Of course, faculty members also benefit from student participation and their frequently innovative input in research projects. Currently, 75 per cent of the papers published by NAU researchers have graduate students as the first author. Recently, one of our PhD students was first author of a paper on jasmonate signalling published in *Nature*.

### **Q. How does NAU contribute to regional and national development?**

The soul of a university lies in its academic spirit, but a university also has the broader mission to serve all of society. This is what we are striving for at NAU. To give an example, a team led by Wan Jianmin researching rice breeding and pest resistance successfully cloned several important genes in rice. The team's research was published in *Nature* and other respected international journals, but more importantly, it helped to control the spread of rice stripe virus in southern China. Another example is NAU's research centre for new rural development, which was founded in line with national priorities for promoting development in rural areas. To foster this type of consequential research, we are now emphasizing a more multidisciplinary approach and stressing social service when evaluating researchers.



Nanjing Agricultural University

# An Agricultural Research Pioneer in China

**As one of the Ministry of Education's '211 project' universities, NAU is dedicated to fundamental and applied research in the field of agriculture.**

A selection of the NAU-led research on critical national and global issues includes:

## Crop Science

Gai Junyi, an academican at the Chinese Academy of Engineering, is dedicated to breeding new soybean varieties and increasing their productivity and quality. Wan Jianmin focuses on breeding rice varieties with disease and pest resistance, the results were published in *Nature*, *Nature Biotechnology* and *Nature Genetics*. Zhang Tianzhen, together with Chen Z. Jeffery at the University of Texas, sequenced the whole upland cotton genome (*Nature Biotechnology*, 2015). Cao Weixing and Zhu Yan's research on crop information technology covers crop growth modelling and general knowledge models for crop management. Ma Zhengqiang focuses on the identification of powdery mildew and scab resistance genes in wheat. Chen Fadi and Hou Xilin study the breeding of chrysanthemums and non-heading Chinese cabbage, respectively, and developed many new varieties. Zhang Shaoling's research uncovered the mechanism of self-incompatibility in pears.

Wu Yidong tested the 'natural refuge strategy for delaying insect resistance to transgenic cotton crops' (*Nature Biotechnology*, 2014). Wang Yuanchao and Zheng Xiaobo

identified new genes in signal transduction and the gene regulation network of plant-oomycete interaction. Han Zhaojun carried out pest resistance target research and established supporting technology, effective in controlling cotton bollworm and *Chilo suppressalis*.

## Animal Health

Lu Chengping has focused on microbiology research associated with livestock. In 2013, Lu's lab was approved by the World Organization for Animal Health as the world's only reference lab for the diagnosis of swine *streptococcus*. Jiang Ping developed two vaccines for swine to effectively control infection with PRRSV and PCV2 viruses. Zhu Weiyun also developed probiotics for swine and enzyme preparations for chickens.

## Food Safety

Zhou Guanhong and Xu Xinglian uncovered the mechanisms underlying the formation of volatile compounds in traditional Chinese cured meat products. They have also established systemic meat grading and quality control standards. Zhou Yingheng focuses on the marketing and circulation of agricultural products, especially relating to issues of food quality and safety risk control.

## Environmental Sciences

Shen Qirong works on technology for the production of bio-organic fertilizers from solid waste material such as straw and animal manure. Zhao Fangjie found that microorganisms in rice paddy soil are able to oxidize and volatilize heavy metals. Pan Genxing focuses on the production and

application of Bio-char from recycled straw. Xu Guohua studies the molecular biology of crop nutrients and water efficient use. Zhang Wenhua and Jiang Mingyi have many achievements in understanding the mechanisms of plant resistance to abiotic stresses.

## Agricultural Economics and Social Science

Zhong Funing and Zhu Jing focus on the theory and policy of national food security in the context of globalization. Qu Futian has introduced innovations in systems of land property rights and regional economic evaluation theory. Wang Siming has been undertaking research on China's agricultural civilization and the Chinese philosophy of science and technology.

Focused on recruiting and retaining high-calibre faculty and building world-class campus facilities, NAU is committed to becoming a leading centre of agricultural education and research.

Nanjing Agricultural University Human Resources: <http://rsrwc.njau.edu.cn/html/en/html/lmy/1.html>

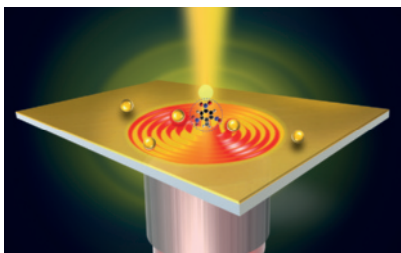


## Contact

Tel:  
Fax:  
Website:

86-25-8439-5754  
86-25-8443-2420  
[www.njau.edu.cn](http://www.njau.edu.cn)





## Shenzhen University

# A rising star in south China

**Shenzhen University (SZU) is a comprehensive research university with state-of-the-art facilities, high-calibre faculty members and a highly professional administration body. Together with the city of Shenzhen — China's most successful Special Economic Zone — the university has been undergoing rapid growth since its foundation in 1983.**

In 2015, SZU received 600 million CNY for research grants and won 205 project grants from the National Natural Science Foundation of China (NSFC). Thomson Reuters' Essential Science Indicators ranks SZU as among the top 1 per cent of institutions in the world for the field of engineering. In 2014, 869 research papers from SZU were published in Science Citation Index journals, including 8 in *Nature* and its sister journals. Some of SZU's recent research is highlighted below.

### Progress in optoelectronics and photonics

Various groups at SZU are conducting groundbreaking research in the areas of optoelectronics and photonics. A group led by Hanben Niu, an academician of the Chinese Academy of Engineering (CAE), made significant contributions to the development of multimode and super-resolution optical imaging. The group developed new form of photodynamic therapy and a non-z-scanning multimolecule fluorescence tracking system with nanometre resolution. These imaging methods have been applied for immunological tracking, three-dimensional DNA imaging and disease diagnosis and therapy and have shed light on life science and clinical research.

A team at the Nanophotonics Research Centre (NRC) led by Xiaocong Yuan has determined how to manipulate arbitrary

focused plasmonic fields to achieve polarization-controlled directional coupling of surface plasmon polaritons. The researchers further proposed and verified focused plasmonic tweezers that can trap and rotate metallic particles or nanowires. Yuan's group has succeeded in manipulating optical vortices in free-space optical communication, achieving high communication capacities. Their research is crucial for the development of next-generation optical communication and interconnection technologies.

Well-aligned, single-chirality, single-walled carbon nanotubes have been used by a team led by Shuangchen Ruan at Shenzhen Key Laboratory of Laser Engineering to manipulate short-pulsed laser states. Being polarization sensitive, these well-aligned nanotubes are ideal saturable absorber materials for the switching of high-power, short-pulsed fibre lasers.

Headed by Dianyan Fan, a CAE academician, the Collaborative Innovation Centre for Optoelectronics Science and Technology is focusing on developing optoelectronic technologies using new advanced materials. Fan's research has contributed greatly to the development of laser-induced fusion in China. A team led by Han Zhang are exploring a new two-dimensional layered material, black phosphorus quantum dots, which are promising for clinical applications. Another team, led by Wenjing Zhang, has reassembled isolated atomically layered two-dimensional materials into hybrid heterostructures, creating new artificial



systems with rich functionalities and novel optical properties.

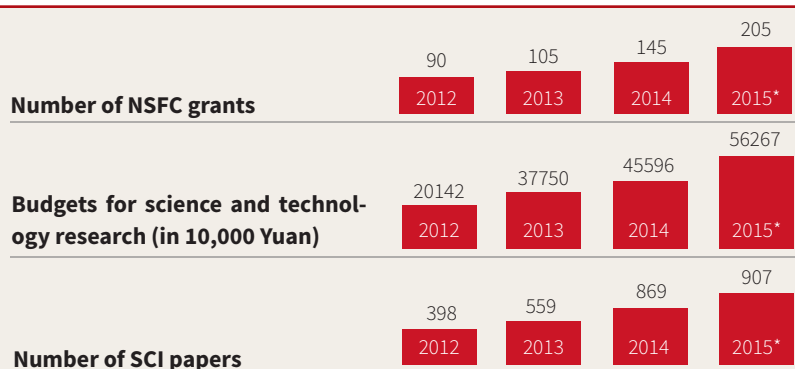
A team led by Yiping Wang, recipient of the National Science Fund for Distinguished Young Scholars of China, is devoted to the design and fabrication of sensing devices in optical fibres to develop all-optical micro total analysis systems. Aiming to create an optical fibre that can act as all-in-one lab, or a 'lab-in-fibre', the researchers are working on microfabrication technology, in-fibre microstructures and novel functional materials.

### Researching engineering and materials

A team led by Qingquan Li, president of Shenzhen University, and Renzhong Guo, a CAE academician, is researching multi-source geoinformation acquisition and services. The team has pioneered new techniques to dynamically acquire spatial data and apply them to geoenvironmental monitoring. The researchers are also exploring new methods for large-scale vehicle and individual trajectory data analysis as well as data mining of social networks. The team won second prize for the National Technology Invention Award of China for their outstanding work on road checking and measurement.

Guoliang Chen, a Chinese Academy of Science (CAS) academician, and his team at the Guangdong Province Key Laboratory of Popular High Performance Computers (PHPCs) have been striving to build high-reliability, low-cost and easy to use PHPCs. They have built KD- and SD-series PHPCs based on the China-made Loongson CPUs. The team has also designed a parallel computing framework that consists of universal representation, partitioning and parallel computing of big data, simultaneously addressing the challenges of volume, velocity and large variety of big data.

A team supervised by Feng Xing developed a service-life design theory of marine structures, involving studies on the failure mechanism for material and structure and the development of novel materials. The team won second prize in the State Technological Innovation Awards as well as two ministerial and provincial-level awards of China.



\* By the end of October, 2015

A group led by Florian Stadler is investigating soft-matter physics and chemistry with a special focus on the rheology of multistimuli polymers and polymer gels. Research on zwitterionic polymer solutions unveiled the influence of ions and zwitterion content on their properties. Polymer blends of dendrimers and polystyrene were found to challenge the foundations of several theories on the phase structures of immiscible polymers.

### Exploring medicine and life science

The functions and mechanisms of selenium and icariin in resisting Alzheimer's disease have been uncovered by a research team led by Jiazuan Ni, an academician of CAS. The researchers have also made remarkable progress in screening biomarkers for early diagnosis of Alzheimer's disease.

A group led by Hong Li is exploring how emotional stimuli affect executive cognitive processing. The group have established that the changes in functional connectivity dynamics are associated with vigilance network, shedding light on the relationship between the dynamics of functional brain networks and individual behaviours by distinct cortical processes.

Yuejia Luo and his team are exploring a broad area of social cognitive neuroscience. They combine brain imaging methods, autonomic measures and behavioural observation methods to investigate mood disorders and understand their neural mechanisms. The research has great clinical

implications for mental disorders and brain diseases.

Functional connectivity among critical language regions in the human brain is being investigated by a team led by Lihai Tan. They have found cross-language differences in the brain networks serving speech and reading, lending support to the culture-specific theory of cortical organization of language.

Xiongzhong Ruan and his group are focusing on lipid-mediated chronic kidney diseases. They have identified a 'wiring diagram' of lipid trafficking under inflammatory stress, demonstrating the mechanism by which inflammatory stress modifies lipid homeostasis. Their study updates the conventional understanding of the pathogenesis of lipid-mediated tissue injury and has important clinical implications.

A research team led by Deming Gou has developed a simple, sensitive and specific method for detecting circulating miRNAs, providing a promising tool for clinically diagnosing diseases based on miRNA biomarkers. They have also identified a group of miRNAs associated with pulmonary arterial hypertension.

### Recruitment of talented researchers

SZU is seeking talented researchers and warmly welcomes outstanding scholars from around the world. We offer excellent compensation packages with large start-up funds and a free intellectual environment. SZU strives to be an open, globally recognized leading university.

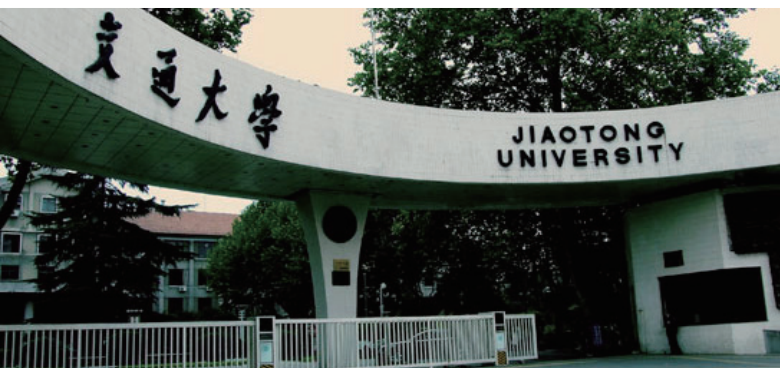
## Contact

**Phone:** +86-755-26534750  
**Website:** [www.szu.edu.cn/2014/en/](http://www.szu.edu.cn/2014/en/)



深圳大学  
SHENZHEN UNIVERSITY





Frontier Institute of Science and Technology

## An international and multi-disciplinary research environment

Frontier Institute of Science and Technology (FIST) was established by Xi'an Jiaotong University in an effort to create a world-class, multi-disciplinary research institute. FIST is the first "special academic zone" at the university. It has introduced an international, scientific research management system that seeks to reform scientific research in China.

Since its establishment, FIST has set up 11 multi-disciplinary research centres, which cover physics, chemistry, biology (including the life sciences and basic medicine), materials science, mathematics, computational science, engineering and other subjects.

FIST aims to drive rapid innovation by becoming a hub for talented researchers from all over the world. So far, 44 scholars have joined FIST, over 40 percent of whom are either academicians or have been awarded national titles. Over the past five years, FIST has grown into a nationally and internationally renowned research institute that has made significant scientific contributions.

## Contact

Tel/Fax: +86 29 83395131  
E-mail: [fist@xjtu.edu.cn](mailto:fist@xjtu.edu.cn)  
Website: [fist.xjtu.edu.cn](http://fist.xjtu.edu.cn)  
Address: 1 West Building, 99 Yanxiang Road, Yanta District, Xi'an, Shaanxi Province, P. R. China, 710054

## That's why I chose FIST!



### Yanzen Zheng

(1000 Young Talent program scholar; principal investigator at the Center for Applied Chemical Research)

After completing Marie Curie Fellowship in 2011, I decided to return to China. FIST immediately drew my attention because, despite being in its infancy, it offered a unique international and inter-disciplinary environment. I believe that this is an ideal model for modern universities and research institutes. Thanks to the great support from Xi'an Jiaotong University over the last four years, both FIST and my research group have grown rapidly. As the first explorers of a new system, we are very proud of the great success of the inter-disciplinary research advocated by FIST. Based in the historical city of Xi'an - the first stop on the former Silk Road - FIST is expected to take a leading role in bringing a new wave of creativity and innovation to China.



### Xiaojie Lou

(1000 Young Talent program scholar; principal investigator at the Multi-disciplinary Materials Research Center)

The 21st century has witnessed China's rapid economic and social growth. As one of only a few inter-disciplinary research institutes in China, FIST provides me with a better environment and more opportunities for doing cutting-edge research than other universities. I enjoy the freedom to pursue my areas of interest and the open-minded atmosphere at FIST.



### Guanghao Lu

(1000 Young Talent program scholar; principal investigator at the Multi-disciplinary Materials Research Center)

FIST is a paradise for academic research. I chose to join FIST because I believe that inter-disciplinary collaboration makes research easier. At FIST, we get to interact with researchers from different disciplines, with different research experience and backgrounds. The laboratories at FIST are well equipped for studying materials science, chemistry, physics and biology. Well known as the eastern terminal of the former Silk Road, the city of Xi'an is located in the Weihe River valley in the heart of China. Xi'an was the capital of China for more than 1,000 years and is home for many famous historical sites, such as the Terracotta Army, City Wall and Wild Goose Pagoda. The beautiful Qinling mountains are only 15 kilometres away.



### Pengfei Li

(principal investigator at Center for Organic Chemistry)

FIST provides me with the freedom to conduct research in whatever subjects I am interested in. It offers very good working conditions, as well as an excellent atmosphere for inter-disciplinary collaboration. The relatively simplified bureaucratic procedures and the support services provided by the administrative staff allowed me to rapidly build my labs and now releases me from the burden of paperwork so that I can spend more time on my research. Furthermore, the unparalleled historical and cultural richness and rapid modernization of Xi'an city makes a very good mix for enriching my life here.